

The international journal of science / 19/26 December 2019

Innovations in
AI and digital health

nature

ONE YEAR. 10 STORIES.

10 people who mattered this year

365 days in science

From the climate strike to quantum computers and beyond

Prehistoric art

Earliest-known example of pictorial storytelling

Events directory

The *Nature* guide to global science events and courses in 2020

Vol. 376, No. 7107
nature.com

The scientific events that shaped the decade

The 2010s saw breakthroughs in fields from gene editing to gravitational waves. The coming decade must focus on climate change.

Scientific and technological innovation has always created social and economic transformation. But the past decade showed, as few others have, the speed and scale at which such change can happen. If it continues at the present rate, the shape of the next ten years – from information technologies to applied bioscience, energy and environment – looks ever more contingent on the discoveries made in that time.

In the 2010s, artificial intelligence (AI) finally began to reveal its remarkable power and disruptive potential. Driven mainly by the advent of deep learning – the use of neural networks to spot patterns in complex data – AI flexed its muscles by achieving reliable language translation, besting expert human players at poker¹, video games² and the board game Go³, and beginning to demonstrate its use in self-driving cars (see *Nature* 518, 20–23; 2015).

Few fields are untouched by the machine-learning revolution, from materials science to drug exploration; quantum physics to medicine. Moreover, it now cannot be doubted that many jobs currently performed by humans could be done more cheaply and efficiently by machines – and the transition might well come sooner than we expect.

These impacts have intensified discussions of risk, but the current danger is not from a Terminator-style robot insurrection. Instead, it will come from inappropriate – or simply bad – uses of the computational tools at our disposal. Algorithms are still unable to automate many human qualities, such as the subtle cognitive capacities we otherwise call common sense. Tomorrow's machines will need to make use of nuanced reasoning and more accurate representations of reality, which demands conceptual advances and architectural innovations as well as bigger circuits.

Appropriate uses of AI must also acknowledge that algorithms trained on the results of past human performance are likely to inherit our biases and prejudices, banishing the idea that an automated process is inherently an objective one⁴. Scientists attempting to develop more humane and dependable AI in the coming decade must heed the call for an interdisciplinary science of 'machine behaviour'⁵ that draws on the skills of psychologists, sociologists, philosophers, legal scholars

and researchers in other disciplines of the social sciences and humanities, along with specialists in engineering and physical sciences.

Regulating at speed

The influence of the information revolution has been felt most strongly in data-rich fields of research. In the life sciences, this has helped to transform the study of the microbiome, the genetic material of all the micro-organisms found in particular environments (see go.nature.com/2yy70bo). In turn, this has affected everything from our appreciation of the importance of microbes in how organic matter decomposes, to our understanding of their role in human diseases. Similarly, the study of human evolution has expanded its focus beyond bones and stones to also include genes and proteins that are now helping to reveal the complexities of evolution, migration and population structure (see go.nature.com/38pdi6m).

It was clear by 2010 that the glut of information made available by the falling costs and growing speed of genome sequencing was going to be both valuable and challenging. But since then, there have been some stark wake-up calls.

Some researchers are deploying Big Data and computing power to explore genetic contributions to highly complex issues, such as behaviour or educational attainment (see *Nature* 574, 618–620; 2019). The reality is that any such links are diffuse and poorly understood. In spite of this, companies offering genetic tests are expanding into what they see as a potentially lucrative market for 'predicting' intelligence, and it is likely that products claiming to predict other traits will follow. This is happening ahead of any consensus among researchers about the reliability and value of such tests, let alone their proper regulation.

Another frontier that researchers have continued to push forward in the past decade is the reprogramming of mature human cells to a stem-cell state. The ability to induce pluripotency – the capacity to transform into multiple tissue types – makes it feasible to grow new cells of almost any variety from adult cells. These are now finding use in exploratory clinical procedures for treating degeneration or damage of retinal and neural tissue – but here, too, there is a burgeoning market for unproven and potentially unsafe 'treatments'.

The 2010s also saw the CRISPR–Cas9 technique^{6,7} harnessed in the cause of genome editing. But for years, the consensus was that no scientist would go so far as to edit a gene in the germ line – human sperm, eggs or embryos – given both the possible dangers to any resulting child, and the unresolved ethical issues involved in making heritable changes. That situation changed, however, when the scientist He Jiankui announced in November 2018 that he had used CRISPR to edit a gene in two baby girls born as a result of *in vitro* fertilization, drawing worldwide condemnation (see *Nature* 563, 607–608; 2018).

As the World Health Organization and academies of science and medicine race to draw up guidelines for regulation, we need to reflect on why ethical and

Few fields of research are untouched by the machine-learning revolution.”



regulatory frameworks have lagged behind scientific and technological advances (see *Nature* 575, 415–416; 2019). At the same time, researchers must consider what can be done now to ensure that technologies are not implemented unless they are shown to be sufficiently safe, effective and inclusive (see *Nature* 576, 7–8; 2019). Here, too, is the troubling possibility that a market demand based on false promises will ride roughshod over the sober deliberations of the scientific community.

Space-time rocks

The 2010s also showed why, in anticipating the coming decade, we should never underestimate scientists' ingenuity and their ability to overcome the odds – given enough resources, and support from funding agencies, industry and decision makers.

In 2008, researchers at CERN, Europe's particle-physics laboratory near Geneva, Switzerland, switched on the Large Hadron Collider, one of the world's most expensive scientific collaborations. In 2012, they confirmed^{8,9} that they had found the Higgs boson, as predicted by particle physics' standard model.

Four years later, in 2016, the announcement that researchers had detected gravitational waves¹⁰ represented the success of a technique that many originally deemed too difficult. The general theory of relativity had long predicted that violent astrophysical events might cause tiny oscillations in space-time; this idea was eventually confirmed by laser interferometry of breathtaking precision. Two experiments – the Laser Interferometer Gravitational-Wave Observatory in the United States and Virgo in Italy – have now been able to measure changes in the dimensions of space-time of a fraction of the diameter of a proton, caused by waves created in the collisions of black holes or neutron stars. With more detectors coming online and powerful upgrades implemented on existing instruments, gravitational waves are now taking their place as windows on the universe, alongside electromagnetic frequencies from radio waves to γ -rays.

Similarly, as the decade began, quantum computing looked like a good idea on paper but a distant prospect in practical terms. Not so today: even the field's specialists have been surprised at how quickly the first devices have evolved. IBM made its five-quantum-bit computer available on the cloud in 2016; the decade ends with machines from IBM, Google and others boasting quantum-bit arrays an order of magnitude larger. One big challenge in the next decade will be to find more ways to make use of these resources, by developing a wider range of quantum algorithms.

China's development of quantum information technologies is just one indication of the nation's remarkable rise as a research superpower. Chinese scientists have used quantum methods to secure long-distance data transmission, for example by pioneering the use of quantum teleportation¹¹ to send information across the world securely by satellite, and by installing an inter-city fibre-optic network that constitutes the first stages of a

If carbon emissions are not drastically reduced by 2030, we will be entering uncharted territory.”

quantum internet. China's government is also looking to shape the global landscape of research in its Belt and Road Initiative: a programme to build infrastructure, including roads, railway lines, ports and even whole cities, all over the world (see *Nature* 569, 5; 2019).

The coming climate crunch

Environmental crises have become depressingly familiar in the past decade, and the alarming rate of global warming lies behind many of them. The latter half of the decade – 2015 to 2019 – was the warmest five years on record, according to the World Meteorological Organization. The pace of warming means that the window for avoiding temperature rises of 1.5 or 2 °C above pre-industrial levels is now frighteningly small. The 2020s will be make-or-break. If carbon emissions are not drastically reduced by 2030, we will be entering uncharted territory, including the possibility – albeit subject to much debate – of passing irreversible tipping points¹², such as the widespread loss of Antarctic ice.

Many countries are now investing for the long run in new energy technologies. The next milestone in the promise of fusion-powered energy will be the switching-on of the international ITER reactor in the south of France in 2025. But any benefits of fusion are too far off considering the urgency of climate change. ITER's road map places the moment of sustainable net power gain around 2035, with commercialization unlikely until at least mid-century.

That means that other ways to create energy while reducing carbon emissions need to become viable on a large scale in the coming ten years. Researchers must pursue innovative technologies such as carbon capture or splitting water through artificial photosynthesis, but solutions must also include significant changes to how the energy economy is run. Navigating a more sustainable path will require ambitious political and industrial will, as much as scientific ingenuity.

In many countries – especially those afflicted by varying degrees of authoritarianism and climate-change denial – that will is in short supply. But researchers must not lose hope. Working with civil society, they must step up, get out of their comfort zones and recognize activism as part of their mission¹³. And they must fight to restore the status of facts and truth.

The 2010s were both remarkable but also troubling. With new knowledge, and a renewed dedication to social and environmental responsibility, the 2020s must be transformational.

1. Brown, N. & Sandholm, T. *Science* **365**, 885–890 (2019).
2. Vinyals, O. et al. *Nature* **575**, 350–354 (2019).
3. Silver, D. et al. *Nature* **550**, 354–359 (2017).
4. Zou, J. & Schiebinger, L. *Nature* **559**, 324–326 (2018).
5. Rahwan, I. et al. *Nature* **568**, 477–486 (2019).
6. Jinek, M. et al. *Science* **337**, 816–821 (2012).
7. Cong, L. et al. *Science* **339**, 819–823 (2013).
8. ATLAS Collaboration et al. *Phys. Lett. B* **716**, 1–29 (2012).
9. CMS Collaboration et al. *Phys. Lett. B* **716**, 30–61 (2012).
10. Abbott, B. P. et al. (LIGO Scientific Collaboration and Virgo Collaboration) *Phys. Rev. Lett.* **116**, 061102 (2016).
11. Ren, J.-G. *Nature* **549**, 70–73 (2017).
12. Lenton, T. M. et al. *Nature* **575**, 592–595 (2019).
13. *Nature* **573**, 309 (2019).

World view

We need a science of the night

Understanding what happens in cities after sunset is crucial to global sustainable development, argues Michele Acuto.

This week, as the Northern Hemisphere has its longest night, my thoughts turn to the 4,300 dark hours in a year. My interest in cities after hours began a decade ago with a project on the economy of waste that saw me riding in rubbish lorries in London, Singapore and Sydney, Australia, often between 11 p.m. and 5 a.m. After being steeped in 'global city' research, spending time with refuse collectors showed me a new world of office cleaners, health-care workers, mammoth restricted-hour lorries, sex workers and homeless people.

This world matters: New York's nightlife accounts for US\$29 billion of economic activity annually and 250,000 jobs. More than one-fifth of Tokyo's workforce is doing night shifts. One-third of everyone employed in London, 1.6 million people, work at night. One-third of transactions in Hong Kong happen after hours. Those who work night shifts face more hazardous and stressful labour conditions than their daytime counterparts.

Yet scholarship and policy often neglect these dark hours, even as research and policy aimed at creating better cities is gaining traction. A more-equitable and sustainable world needs a science of the night, an explicit cross-disciplinary focus to inform policy about the issues faced by urban areas at night, from energy to climate, waste and inequality.

Some cities are waking up to this. In 2018, New York, the 'city that never sleeps', set up an Office of Nightlife. In 2012, Amsterdam appointed its Night Mayor. London has had a Night Czar since 2016. These scattered efforts need to be better coordinated and connected. And there is still too little rigorous evidence to inform policy.

For instance, few analyses look to see whether policies exacerbate inequalities, which tend to be worse at night. The hospitality and entertainment sectors get most of the focus, even though more midnight workers are employed in logistics and health care. Work at University College London (UCL) demonstrated that night-time spaces for LGBT+ people (people from sexual and gender minorities) are important for community life, and are also at a higher risk of closing than other establishments (B. Campkin and L. Marshall *Soundings* 70, 82–96; 2018). UCL also highlighted inequality in transport options: London's celebrated 24-hour Night Tube serves bustling downtown and restaurant districts, and so does more to accommodate late-night revellers than low-income late-shift workers.

In 2018, at least 8,855 people slept rough on the streets of London, a 140% increase over the past decade, with similar trends globally. An estimated 154 million people, about

Some of the most influential thinking on the night-time comes from fast-food chains."

Michele Acuto

is a professor and director of the Connected Cities Lab at the University of Melbourne.
e-mail: michele.acuto@unimelb.edu.au



By Michele Acuto

2% of the world's population, are homeless. Nights for them are often physically and mentally dangerous. A 2015 study found that the health effects of heatwaves at night, including on sleep and stress levels, have been overlooked in Australian cities, and probably on other continents.

Information about the night-time is also crucial for a sustainable planet. At the Connected Cities Lab, we are working with the Melbourne School of Design and the London-based design firm Arup to evaluate how cities are performing at night-time vis-à-vis the United Nations Sustainable Development Goals. This is no academic exercise. Evidence that late-night and shift workers have higher risks of conditions such as heart disease, mental-health disorders and cancer reinforce other analyses calling for a higher night-time wage. Understanding that energy use often peaks at night calls for a smarter lighting infrastructure across our cities. Appreciating effects on wildlife can encourage 'temporal zoning' that benefits plants and animals in our cities. Our goal is to provide evidence on the complexity and value of this knowledge, while preparing the next generation of city leaders to be savvier about night-time risks.

A science of the night should build on many existing pockets of promise within and beyond academia. After-hours analytics in neuroscience and physiology is thriving. The 'science of sleep' is ever-popular in the media, and bio-science about nocturnal changes in animal behaviour is making headway. In the humanities, literary and cultural studies of 'nightwalking' have a long tradition of depicting how our society changes after hours. The British geographer Robert Shaw has even called for the development of a 'nightology' in the social sciences.

Night-oriented research does not yet form a coherent body of inquiry, and there is too little discussion about how night-time shapes sustainability challenges worldwide. Social and natural sciences need to speak to each other more about what happens after hours. Especially when some of the most influential thinking on the night-time comes from fast-food chains, utilities giants and other industries that look to maximize sales without considering effects on places and people. McDonald's, as a \$30-billion real-estate powerhouse with an explicit night-time strategy, probably knows more about the after-hours than most policymakers.

Night-time literacy programmes for scientists and officials are urgently needed. These should deliver better night policy, from zoning to wages, transport and nature-based solutions. Night-time assessments beyond the entertainment sector should be expected when mayors promote action on global issues such as climate, resilience or migration. Night-time is a globally shared experience across countries and cultures. By building local knowledge, we can craft global good. It's time to bring what happens when the lights go down out of the shadows.

News in brief

GLOBAL PROJECTS SPARK SURGE IN THOUSAND-AUTHOR PAPERS

The number of papers with over 1,000 authors has more than doubled in the past 5 years, a study of millions of articles indexed by the Web of Science (WoS) database has found.

Between 2009 and 2013, 573 manuscripts were published with 1,000 co-authors or more, according to a 4 December report by the Institute for Scientific Information (ISI), part of Clarivate Analytics in Philadelphia, Pennsylvania, which runs the WoS. But that figure has risen to 1,315 papers over the past 5 years.

The surge in this practice, dubbed hyperauthorship, reflects the increasingly global nature of research in several fields, the institute says.

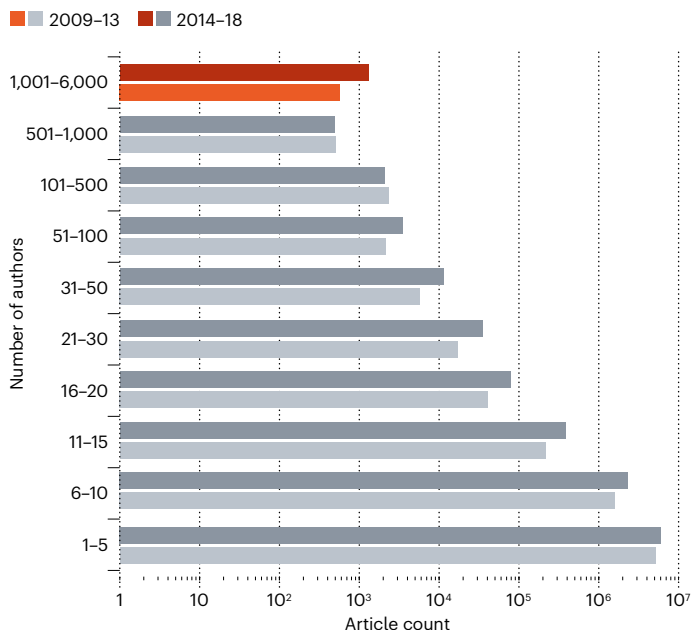
In particle and nuclear physics, papers with hundreds or thousands of authors have been common for some time, mostly because of massive collaborative research projects at CERN, Europe's particle

physics laboratory near Geneva, Switzerland. But data from the past five years show an increasing number of thousand-author studies in other fields, including global epidemiology and climate change, says Martin Szomszor, head of research analytics at the ISI. A 2017 analysis of body-mass index published in *The Lancet*, for example, involved more than 1,000 authors from more than 100 countries (NCD Risk Factor Collaboration (NCD-RisC) *Lancet* 390, 2627–2642; 2017).

Szomszor thinks the trend will continue over the next five years, as more fields start to involve collaborations between large numbers of researchers in different parts of the world. He predicts that topics associated with the United Nations Sustainable Development Goals — such as poverty and sustainability — are likely to produce more mass-authored papers in future.

HYPERAUTHORSHIP

In recent years there has been a significant increase in the number of papers with more than 1,000 authors.



ROGUE STEM-CELL SALESMAN DIES

The disgraced stem-cell entrepreneur Davide Vannoni, who in the past decade treated hundreds of people in Italy with an unproven treatment that health authorities considered dangerous, died on 10 December after an illness. He was 53.

His death brings to an end a years-long saga in which a group of academic stem-cell researchers in Italy fought to halt his activities, which they said endangered people.

Vannoni was not a trained scientist or physician but in 2007 he claimed to have invented a stem-cell therapy that could cure neurological conditions, such as Parkinson's disease and muscular dystrophy. The procedure, he said, converted stem cells taken from a person's bone marrow into neuronal stem cells, which were then infused back into that person. He sold the therapy through his organization, the Stamina Foundation.

In 2012, the Italian Medicines Agency prohibited the treatment. But Vannoni had passionate supporters who often mounted legal challenges to allow them to try the 'Stamina method'.

Prosecutors in Italy fought to ban the procedure, and in 2015 courts in Turin convicted Vannoni on fraud charges. He was later jailed, but was released last year because of his ill health.



New Zealand volcano eruption 'worst of all possible scenarios'



At 2:11 p.m. in New Zealand on 9 December, an explosive eruption forced superheated steam, boulders, caustic ash and other debris out of the White Island volcano, creating a 3.7-kilometre-high column of ash. So far, 16 people have been confirmed dead, and many others remain in hospital with severe burns.

Geologists who had been monitoring the marine volcano in the months preceding the eruption had seen an increase in volcanic activity, but nothing that could have immediately predicted the disaster.

The event was the worst of all possible scenarios, says Raymond Cas, a volcanologist at Monash University in Melbourne, Australia. Rocks 1 metre or more across would have been ejected as ballistic blasts, he explains. “Finer debris would have created an ash cloud that reduced visibility to zero, making it impossible for people to know where to run to.”

White Island sits off the east coast of the country’s North Island. It is one of the country’s most active volcanoes. The explosion was either a hydrothermal or a ‘phreatic’ eruption, both of which are caused by the build-up of pressure of superheated steam and gas.

SCIENTISTS WALK FREE IN US-IRAN PRISONER SWAP

The US authorities have dropped charges against two researchers accused of attempting to export chemicals to Iran in violation of trade sanctions. The decision follows a prisoner swap in which the United States freed a researcher who was charged alongside them – and imprisoned. Iran released a jailed US researcher in exchange.

The scientist freed by the United States is Masoud Soleimani (pictured, below right), an Iranian biologist arrested last October as he was about to take up a research position at Mayo Clinic in Rochester, Minnesota. The same day, Iran’s government released historian Xiyue Wang, a US citizen jailed in Iran in August 2016 on spying charges.

US authorities have also dropped charges against two of Soleimani’s former students, Mahboobe Ghaedi and Maryam Jazayeri, both of whom are in the United States.

Iranian authorities had imprisoned Wang, who is a PhD candidate in Persian history at Princeton University in New Jersey, in 2016. He had been working in Iran’s national archives in Tehran.

Brian Hook, the US special representative on Iran, said that the exchange is “a good first step, and I hope this leads to bigger and better things”.



CHINESE LABS INVESTIGATE PATHOGEN OUTBREAK

Two Chinese institutions are investigating how more than 100 students and staff became infected with the bacterium *Brucella*, strains of which are typically found in farm animals but can cause severe illness in people.

The Lanzhou Veterinary Research Institute in central China confirmed on 7 December that 96 staff and students have tested positive for the infection. It said most of the infected people are not experiencing signs of brucellosis, the illness caused by the bacterium, which can include fever and flu-like symptoms. The institute says it has begun an investigation.

On 10 December, the health commission for Heilongjiang province confirmed that 13 students at the Harbin Veterinary Research Institute also had the infection. The 13 were among 49 students who had previously worked as interns at the Lanzhou institute.

Different strains of *Brucella* occur in many mammals, including goats, sheep, cattle and pigs. Human infections most commonly result from the ingestion of undercooked meat or raw milk – but the bacteria can also enter the body through the lungs or skin wounds. *Brucella* is not typically transmitted between people. If left untreated, the infection can travel to the heart or brain and, in rare cases, be fatal.

News in focus



HOLLIE ADAMS/BLOOMBERG/GETTY

Boris Johnson remains UK prime minister following a Conservative victory in the UK general election.

UK ELECTION DASHES SCIENTISTS' HOPES OF STAYING IN THE EU

Brexit certainty intensifies researchers' concerns about international collaboration and access to European Union funds.

By Jonathan O'Callaghan

The United Kingdom is now firmly on the path towards leaving the European Union, after the Conservative party won a majority of 79 seats in last Thursday's general election – a result that has major implications for science.

In the wake of the election, researchers are also questioning whether the party will be able to honour its campaign promise to increase spending on science.

Prime Minister Boris Johnson campaigned on the basis that he would take the United

Kingdom out of the EU with his previously negotiated withdrawal agreement by 31 January 2020, if his party won a majority. So the Conservative victory more or less ends the possibility of remaining in the EU, an outcome that was left open before the general election, and which some scientists had hoped for.

"Given the pro-remain sentiments of a large majority of the scientific and academic community, many people would have been clinging to the hope of some kind of second referendum or some attempt to try and reopen the fundamental question," says James Wilsdon, director of the Research on Research Institute

at the University of Sheffield, UK. "Clearly, that option has now gone."

But the result does mean that, for the time being, researchers no longer face the prospect of a chaotic no-deal Brexit.

"There is a great degree of certainty in what was a very uncertain situation," says Wilsdon. "I think a lot of the science community don't like the substance of that certainty, but at least this does mean we won't be looking at months or years more of when or how we'll be leaving the European Union."

Although Brexit now looks certain, there is still a question mark over what the United

Kingdom's future relationship with the EU will look like.

Details on trade and other aspects have yet to be ironed out, while key issues for science – such as the United Kingdom's involvement in Europe's Horizon 2020 research programme, a crucial source of funding and collaboration – have yet to be resolved. “The Conservative manifesto says we will continue to collaborate internationally and with the EU on scientific research, including Horizon,” says Sarah Main, executive director for the Campaign for Science and Engineering in London. “But it's not quite 100% clear how that's going to be enabled to happen.”

Brexit will also bring changes to the free movement of EU citizens in and out of the United Kingdom, which could affect overseas recruitment at UK universities and research institutions. The Conservatives promised in their manifesto to introduce “new rules for those of exceptional talent” in a post-Brexit immigration system.

It's now necessary to ensure that non-British, European researchers who currently benefit from freedom of movement can still come to the United Kingdom, says Beth Thompson, head of UK and EU policy at Wellcome, a biomedical-research charity in London. “It's important that we send a signal to the rest of the world that the UK is open for business, and that we want to participate in internationally competitive and collaborative science.”

Manifesto pledges

Whether the government can fulfil the science promises laid out in the Conservative manifesto is also unknown. The party has committed to raising UK spending on science and research to 2.4% of gross domestic product (GDP) by 2027, up from 1.7%.

But the Conservatives have so far failed to make much progress towards this target, warns Kieron Flanagan, a science-policy researcher at the University of Manchester, UK. The pledge to increase research spending to 2.4% of GDP was made in the run-up to the 2017 general election. “It's been an objective for a few years now,” says Flanagan, “But we haven't seen much activity.” He adds that roughly two-thirds of research funding currently comes from the private sector, so both private and public spending increases will be needed to reach the 2.4% target.

Thompson says that the Conservative manifesto has some “very strong commitments to science”, but at the moment we “don't have detail on how that will be implemented”.

Other Conservative pledges will also come under scrutiny, such as the proposal to develop “a new agency for high-risk, high-payoff research”, thought to be modelled on the US Defense Advanced Research Projects Agency. At the moment, it is still unclear how the agency would actually operate and how it

would improve science in the United Kingdom. “We can all rally around those aims,” says Wilsdon. “But I've not seen anything yet that makes a clear, evidence-informed case for why we need a new institution.”

As the new government settles in, researchers will have to wait and see whether

the ruling party can fulfil its manifesto pledges, and how negotiations with the EU progress. “We've got a government that is driving an aggressive and ambitious science agenda, but it also has a mandate to leave the EU,” says Main. “And that raises questions for the science community.”

CHINA SPENDS MILLIONS TO BOOST HOME-GROWN JOURNALS

US\$29-million investment aims to boost the country's status as an international scientific powerhouse.

By David Cyranoski

China is taking dramatic steps to improve the quality and international reputation of its home-grown science journals. Publishers of hundreds of Chinese titles will receive generous government funding as part of a major five-year plan to elevate the country's publications to among the world's best.

The government said in August that it wants to publish more of the world's breakthrough discoveries in Chinese journals. On 25 November, it revealed that it will spend more than 200 million yuan (US\$29 million) per year for 5 years to help improve the standards of some 280 journals – most of which publish

“There is no such thing as Chinese chemistry, American biology or German physics.”

in English – and to increase submissions from international researchers.

China has launched several initiatives over the past 5 years to improve the quality and international submission rates of its roughly 500 English-language science journals, following growing concerns that some were publishing a lot of low quality, even fraudulent, research. The initiatives have helped to improve some publications, but editors say that few manuscripts are submitted from top researchers in China or abroad.

The latest initiative is the largest and most comprehensive attempt yet to transform the country's scientific-publishing landscape, says Tao Tao, an independent consultant on Chinese academic publications who is based in Washington DC. “The new programme, given

its scale, will be successful,” she says.

It also marks a turning point in a long-running debate about how China should raise its status as an international scientific powerhouse, says Tang Li, who researches science policy at Fudan University in Shanghai. Many Chinese-born scientists who have returned after training overseas think the country's research heft is already reflected in the increasing number of Chinese scientists publishing in prominent foreign-owned journals. But Chinese journal editors and publishers think that more highly regarded domestically owned publications are needed to burnish the country's reputation. The latest investment signals that the government is backing the latter strategy, says Tang.

The investment is being overseen by a committee of representatives from seven high-profile organizations: the finance, science and education ministries; the General Administration of Press and Publication, a powerful Communist Party propaganda agency; the Chinese science and engineering academies; and the Chinese Association for Science and Technology, a non-governmental science organization.

To determine how funds will be allocated, the committee has ranked 250 journals into 3 tiers on the basis of quality, although it has not released details about how the ranks were decided. Twenty-two ‘tier one’ journals, which publish in English, will each receive between 1 million and 5.2 million yuan per year to help them attract submissions from researchers around the world. Another 29 ‘tier two’ English-language journals will each receive between 600,000 and 1 million yuan per year. Four hundred thousand yuan will be invested in each of another 199 ‘tier three’ journals, half of which publish in Chinese. An additional 30 journals that have not been ranked will be selected each year to share 500,000 yuan to

improve their quality.

The government has not yet revealed how the programme's success will be measured, but Tao thinks that journal impact factors might be used to gauge improving quality.

The investment is understandable, given that publications don't have a lot of money to boost quality themselves, says Cao Cong, a science-policy researcher at the University of Nottingham Ningbo China. But he notes that the country has mostly succeeded in becoming a research powerhouse without such publications. Science is international and researchers want to publish in the best

journals regardless of where they're based, says Cao. "There is no such thing as Chinese chemistry, American biology or German physics," he says.

Cao doubts that the investment in Chinese-language journals will lead to international acclaim, because non-Chinese-speaking scientists are unlikely to publish in them.

But having more Chinese-owned publications could save Chinese institutions money, says Tao, because – unlike international journals – domestic publishers are likely to offer reduced publication charges for Chinese researchers, she says.

EARTH'S MAGNETIC FIELD IS OLDER THAN SCIENTISTS THOUGHT

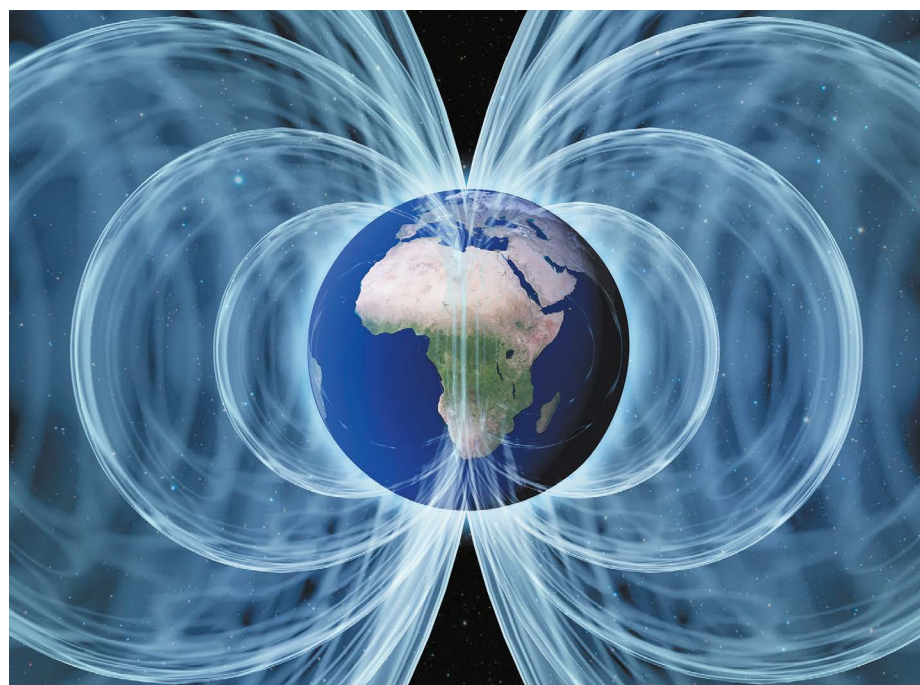
Analysis finds planet's protective shield was in place by at least 3.7 billion years ago, as early life arose.

By Alexandra Witze,
San Francisco, California

Magnetic minerals in ancient Greenlandic rocks suggest that Earth's magnetic field arose at least 3.7 billion years ago. The finding pushes back the time of the magnetic field's birth to about 200 million years

earlier than the commonly accepted estimate – around the time life first appeared on Earth.

Scientists think that having a magnetic field makes Earth more hospitable to life. The field, which is generated by liquid iron sloshing about in the planet's core, shields Earth from energetic particles flowing from the Sun. It helps the planet hold on to its atmosphere and maintain liquid water on its surface.



Earth's magnetic field, shown here as white lines, helps the planet hold on to its atmosphere.

But very few rocks that are billions of years old, and thus could preserve evidence of when the magnetic field arose, have survived to the present day. The new report is a rare glimpse at what Earth was like billions of years ago.

"I hope you are as excited as I am," Claire Nichols, a palaeomagnetist at the Massachusetts Institute of Technology in Cambridge, told a meeting of the American Geophysical Union in San Francisco, California, on 9 December.

Rare rocks

Nichols led two expeditions to western Greenland in the summers of 2018 and 2019. She was targeting a set of ancient rocks in the Isua region, north of the capital city Nuuk, that researchers have long studied in search of clues to early life. The Isua rocks have inspired fierce debates, including whether they contain fossils of complex organisms from 3.7 billion years ago.

Geological forces have squeezed and heated the rocks so much over the past few billion years that most scientists thought the rocks had lost most of their magnetism. But Nichols and her team travelled to the northernmost part of Isua to study rocks that had been least affected by this squeezing and heating.

Iron minerals in those rocks yielded information on the direction of Earth's magnetic field when the minerals formed. Because the rocks are 3.7 billion years old, the magnetic signal must be, too, Nichols said.

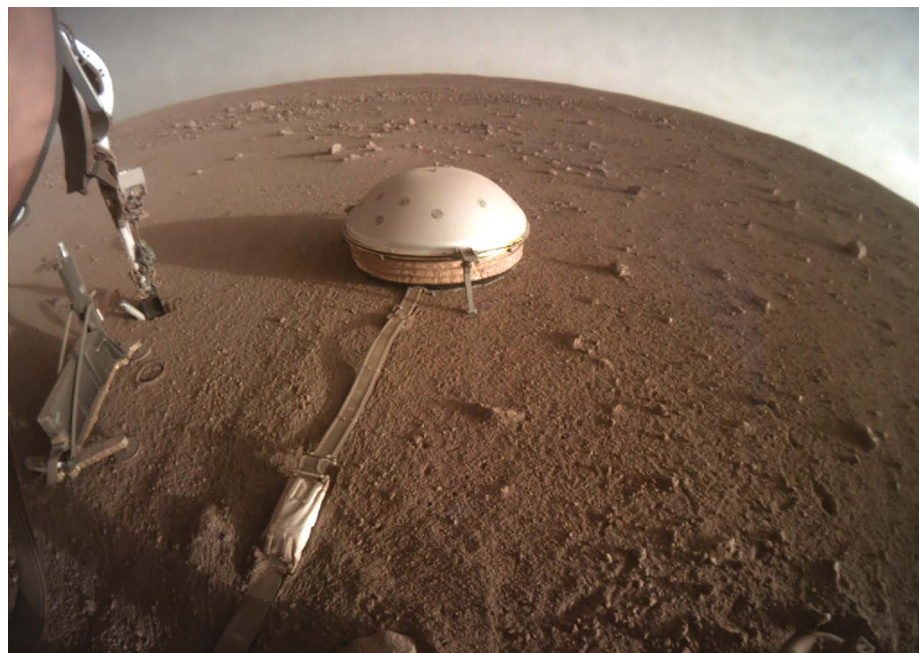
Her team ran various tests to try to confirm that the signal was real and not some sort of weak magnetism introduced later as the rocks were heated and squeezed.

"It does sound super-exciting," says Nicholas Swanson-Hysell, a geoscientist at the University of California, Berkeley, who was in the audience at Nichols's talk. He met up with her afterwards to brainstorm ideas about how to confirm her team's finding. One idea might be to look at rocks from parts of north-eastern North America that were connected to Greenland in the past, to see whether they can illuminate more of the geological history of the Isua rocks, he says.

John Tarduno, a palaeomagnetist at the University of Rochester in New York, is more sceptical of Nichols's claim. "I'd like it to be true, but I'd like to see more," he says.

In 2015, Tarduno and his colleagues reported finding signs of Earth's magnetic field from more than 4 billion years ago, inside zircon crystals from Australia. Other scientists recently challenged that paper, saying the magnetic minerals inside the zircons could not be accurately dated (F. Tang *et al. Proc. Natl Acad. Sci. USA* **116**, 407–412; 2019).

Aside from those contested Australian zircons, the oldest-known evidence of Earth's magnetic field – rocks in South Africa – dates to around 3.5 billion years ago.



The Mars lander's seismometer contains three sensitive sensors nestled inside a dome.

'MARSQUAKES' REVEAL RED PLANET'S HIDDEN GEOLOGY

NASA's Mars InSight lander has detected more than 300 quakes and traced some back to their source.

By Alexandra Witze,
San Francisco, California

The marsquakes are coming fast and furious. From its landing site near the Martian equator, NASA's InSight mission is detecting about two quakes per day – and the rate is going up.

"We have a lot," said Bruce Banerdt, a geophysicist at the Jet Propulsion Laboratory in Pasadena, California, and InSight's principal investigator. He reported the findings on 12 December at a meeting of the American Geophysical Union in San Francisco, California.

Since arriving on Mars just over a year ago, InSight has detected 322 marsquakes. They are the first quakes ever detected on any body other than Earth or the Moon. Scientists aim to use them to probe the Martian interior, including mapping how the planet's guts are divided into layers of crust, mantle and core.

Most of the marsquakes are tiny, much smaller than anything that would be felt on Earth. But some have been big enough – up to nearly magnitude 4 – for scientists to be able to trace them back to their source.

Two of the biggest marsquakes came from a geologically active area known as Cerberus

Fossae, which lies about 1,600 kilometres east of InSight. The quakes there might have been caused by the build-up of stress along geological faults in the Martian crust, and then released in a marsquake.

Other early findings from the mission include mysterious magnetic pulses that appear at about midnight each night around the lander. But one of InSight's main goals – to hammer a heat probe 5 metres into the Martian ground – remains frustratingly out of reach. The probe, dubbed 'the mole', has encountered more friction in the soil than scientists had expected. In October, it even unexpectedly backed out of its hole.

The biggest discoveries so far have come from the ever-expanding catalogue of marsquakes. InSight's highly sensitive seismometer hunts for quakes at night, after the winds that shake the ground during the day die down.

The marsquakes come in two types. The most common shakes the ground at high frequencies. Less common is a type that is detectable at lower frequencies. The high-frequency signals might be coming from quakes that rupture the shallow Martian crust, whereas the low-frequency ones might be travelling from deeper within the planet, in its mantle, said

Domenico Giardini, a seismologist at the Swiss Federal Institute of Technology in Zurich.

Two of the biggest marsquakes hit in May and July. Both were of the low-frequency type. Team members were able to trace the seismic energy back to Cerberus Fossae. This area is home to recent geological activity, including faults that seem to have moved in the past ten million years.

Before InSight launched, researchers had predicted it might be able to detect quakes coming from Cerberus Fossae. The faults there could build up stress at their ends, said Alice Jacob, a planetary scientist at the Paris Institute of Earth Physics. An analysis she led suggests that this could be the source of the marsquakes picked up by InSight.

The rate of quakes has been increasing, Banerdt said – from a few sporadic tremors reported after InSight landed, to the current pace of two a day. Mission scientists aren't sure why.

Equally mysterious are the magnetic pulses that show up every night. InSight measured them with its magnetometer, and they are thought to be related to something happening in the space environment around Mars. One idea is that they are created when charged particles from the solar wind slam into Mars.

Probe problems

InSight's greatest drama so far has come with its mole. This initially began burrowing into the ground as planned, but hit disaster in October, when it suddenly squirted out of its hole.

Mission engineers designed the mole to work in a type of soil different from that it actually encountered. It was designed for cohesionless soil, in which particles flow with little to no friction between them – as in a vat of sugar. But InSight's landing place turned out to have cohesive soil, in which the particles stick together, more like those in a vat of flour, says Tilman Spohn, a space scientist at the German Aerospace Center in Cologne.

When the mole began burrowing, the soil around it became compacted into a pit. The mole could not build up enough friction against the pit's walls to keep moving into the ground. Spohn says that he and his colleagues had seen this happen in laboratory experiments involving cohesive soils, but that they expected InSight's landing place to have cohesionless soils, in common with other Martian landing sites.

Mission engineers have been trying to get past the problem by pinning the mole to the side of the pit with the lander's arm, to give it more friction to keep going. And it is starting to bury itself into the ground again, slowly and carefully.

"By Christmas time, maybe our present will be that we're back to square one," says Spohn. "Which at this point in time would be a very, very welcome situation."

NASA/JPL-CALTECH



VICTOR MORIYAMA/NYT/REDUX/EYEVINE

Large areas of Brazil's Amazon rainforest were burnt to clear space for cattle ranching.

2019 IN REVIEW

A year marked by climate protests, political uncertainty and debate over the ethics of gene editing in human embryos proved challenging for science. But researchers also celebrated some exciting firsts – a quantum computer that can outperform its classical counterparts, a photo of a black hole and samples gathered from an asteroid.

This year, astronomers glimpsed the blackness of a black hole for the first time ever. In April, the international Event Horizon Telescope collaboration unveiled perhaps the most memorable picture of 2019: the first direct image of a black hole and its event horizon (see page 354). To produce it, researchers coaxed a network of radio telescopes to take simultaneous readings from around Earth.

In a year that marked the 50th anniversary of the Apollo Moon landings, lunar exploration was high on the agendas of space agencies. In January, China's Chang'e-4 probe became

the first spacecraft to land safely on the lunar far side. Its rover, Yutu-2, continues to roll across the dusty soils of Von Kármán crater. Other attempts to explore the Moon were not so successful. In April, an Israeli-led effort to put the first private spacecraft onto the Moon ended in a crash landing. The same thing happened to India's Vikram lander in September, although the orbiting part of that mission – known as Chandrayaan-2 – is still circling the Moon as planned.

Ongoing Mars missions returned a host of results. The French-built seismometer on NASA's InSight lander detected the first-ever

'marsquakes'. Roughly 600 kilometres away, NASA's Curiosity rover sniffed record-high levels of methane gas in the Martian atmosphere in June – a mystery that scientists have yet to explain, especially because the methane vanished in days. In February, NASA officially bid farewell to its most stalwart Mars rover, Opportunity.

In the farther reaches of the Solar System, Japan's Hayabusa2 probe collected a sample from the surface of the asteroid Ryugu in February. Then, in July, it dropped a small pellet onto the asteroid and blasted its surface, before descending to gather some of

the freshly exposed material. Hayabusa2 will return its samples to Earth next year. Far beyond Pluto, NASA's New Horizons spacecraft passed a 35-kilometre-long object known as Arrokoth. Its bizarre shape, resembling two pancakes stuck together, gave humanity our closest glimpse yet at an icy, primordial world.

This year even brought a visitor from beyond the Solar System. The interstellar Comet 2I/Borisov whizzed past the Sun earlier this month. It is only the second object known to have visited our Solar System from another one, following 2017's 'Oumuamua.

Heated debate

Back on Earth, it was another tough year for the environment. Up to one million plant and animal species now face extinction owing to habitat destruction and climate change, warned a report by the Intergovernmental Science–Policy Platform on Biodiversity and Ecosystem Services, a panel backed by the United Nations. And the Intergovernmental Panel on Climate Change (IPCC) called for drastic efforts to curb demand for agricultural land, including people shifting towards a plant-based diet, in a special report. Without such action, the IPCC said, governments will fall short of their collective goals under the 2015 Paris climate accord, in which nations agreed to limit global warming to no more than 2 °C above pre-industrial levels.

But political trends seemed to move in the opposite direction. In Brazil, the populist President Jair Bolsonaro took the helm in January with a fiery anti-environmental agenda. He slashed federal funding for science, and in July accused his own government's scientists of lying about a spike in deforestation in the Amazon. In the United States, President Donald Trump continued his efforts to dismantle environmental regulations. In June, the US Environmental Protection Agency (EPA) finalized a plan to relax limits on greenhouse-gas emissions from power plants. In August, the EPA followed up with a proposal to freeze fuel-efficiency standards for automobiles, and the president announced in September that the agency would revoke a long-standing waiver allowing California to set its own limits on carbon emissions from cars and trucks. And in November, the administration began the official process of pulling the United States out of the Paris agreement.

Activists around the world responded to government intransigence with protests, including September's Global Climate Strike. Galvanized by youth climate activist Greta Thunberg (see page 372), millions of people in 150 countries took to the streets to demand stronger action. In October, youth leaders filed a pair of lawsuits against the state of Alaska and Canada's federal government, arguing that they are violating their rights by encouraging the use of fossil fuels. The lawsuits are part of a larger trend in climate litigation, including



Climate-change protestors gather at a demonstration in Cape Town, South Africa.

a major case pending in the Netherlands. In May, the Dutch supreme court heard the government's appeal against a lawsuit brought by the Urgenda Foundation, a citizens' climate organization that successfully argued in the lower courts that the Dutch government must do more to combat climate change. If the supreme court rules in Urgenda's favour, the government will be unable to appeal further.

Pushing biological boundaries

It was a year of testing biological and ethical limits in the laboratory. US researchers revived the brains of pigs four hours after their heads had been severed, by pumping in a nutrient- and oxygen-rich liquid to mimic blood (see page 365). The trick triggered sugar consumption and other metabolic functions, suggesting that the brains were still working. The researchers did not attempt to restore consciousness, however – they added chemicals to stop neurons

from firing before the experiments started.

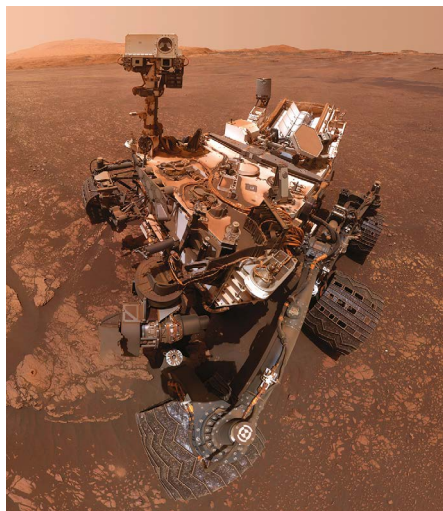
In another out-of-body experiment, scientists grew monkey embryos in a dish for nearly three weeks – longer than primate embryos have ever been grown in the laboratory before. The feat raises the question of whether lab-grown human embryos should be allowed to develop beyond 14 days, a restriction imposed in most countries. In September, a US research team provided a possible circumvention of the 14-day limit by growing a human embryo from stem cells. The 'artificial embryo' seemed to mimic the early development of a real human embryo. Whether it should be permissible to grow artificial embryos to later stages is an ongoing ethical debate.

Japan continued its dominance in the clinical use of induced pluripotent stem cells – adult cells that are reprogrammed into an embryonic-like state. In September, a Japanese group used these stem cells to make sheets of corneal cells that could be transplanted into a woman whose eyesight was failing. In the past decade, Japanese physicians have used iPS cells to treat Parkinson's disease and another eye condition, and this year a group was granted approval to use them cells as a therapy for spinal-cord injury. However, the jury is still out on whether any of these treatments is effective.

Culture shock

Investigations into harassment and workplace culture and ethics have continued at research institutions around the world. Staff at Germany's Max Planck Society reported that gender-based discrimination and bullying are regular occurrences, in a massive employee survey that drew more than 9,000 responses.

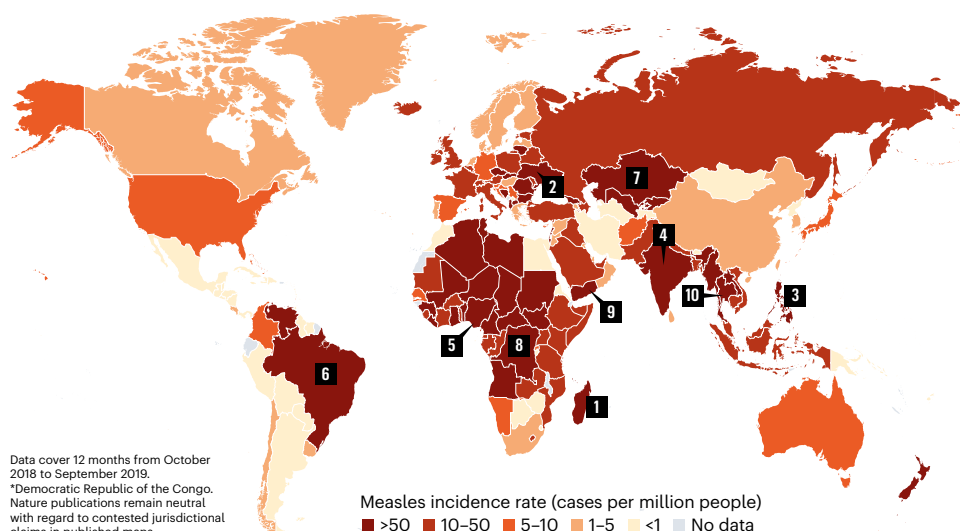
In Australia, 50% of female scientists who responded to a national poll said they faced harassment at work. And in August, the University of Adelaide suspended Alan Cooper,



NASA's Curiosity continued to explore Mars.

MEASLES ON THE RISE

The number of measles cases being reported to the World Health Organization continued to climb in 2019. Provisional data released in November show that severe outbreaks are ongoing in several countries, including Madagascar, Ukraine, the Philippines and Brazil.



Top 10 countries by reported cases

1. Madagascar	151,032
2. Ukraine	78,708
3. Philippines	49,419
4. India	36,251
5. Nigeria	27,954
6. Brazil	18,927
7. Kazakhstan	10,696
8. DRC*	9,245
9. Yemen	9,156
10. Thailand	7,738

SOURCE: WHO

head of the prestigious Australian Centre for Ancient DNA, after an inquiry into the 'culture' at the centre, and amid accusations from some co-workers that he had bullied them. The university has not given a reason for the suspension and told *Nature* that he is still suspended and that "the process in relation to him remains underway".

In the United States, the National Institutes of Health announced for the first time how many of its grant recipients had been disciplined as a result of sexual-harassment investigations in the previous year. The agency said in February that it had replaced 14 principal investigators in 2018, and banned 14 people from participating in its peer-review panels. Meanwhile, the US National Academy of Sciences approved a policy to expel members found guilty of sexual harassment. And the Massachusetts Institute of Technology (MIT) in Cambridge is investigating its links to disgraced and deceased financier Jeffrey Epstein. Epstein had donated about US\$800,000 to the university, MIT says.

Quantum wonders

Physicists reached a long-awaited milestone in quantum computing. In October, a team at Google reported in *Nature* that it had used a quantum computer to perform a calculation that would be practically impossible for a classical machine, even a state-of-the-art supercomputer. The calculation itself – checking the outputs from a quantum random-number generator – is of limited practical use, but the feat is a step towards future applications of quantum computers, which range from designing new materials to codebreaking.

Another Google unit, the London-based artificial-intelligence (AI) powerhouse DeepMind, made headlines when it showed that its programs had mastered the multiplayer online videogame *StarCraft II*. And for the first

time, an AI bot beat human champions at multiplayer poker. Although AIs that can beat the best human players at chess or Go – as DeepMind's AlphaGo did in 2016 – are impressive, many in the field consider multiplayer games to be better analogues of real-life challenges for machine learning, such as fraud detection or self-driving cars.

Earlier in the year, molecular-scale transistors came into view when chemists made the first-ever ring-shaped molecule of pure carbon by using an atomic-force microscope to manipulate individual molecules.

Embryo edits

As 2019 began, the world was still reeling from the announcement that Chinese scientist He Jiankui had produced the world's first gene-edited babies. He used the CRISPR–Cas9 system to alter the gene *CCR5*, which encodes a protein that HIV uses to enter cells, in an attempt to give twin girls resistance to the

virus. In January, the Southern University of Science and Technology in Shenzhen fired He, after a Chinese health-ministry probe found that he had violated national regulations forbidding the use of gene editing for reproductive purposes. In March, the health ministry issued further draft regulations that included severe penalties for those who break rules regarding gene editing in humans. That month, an advisory committee to the World Health Organization called for the creation of a global registry of human gene-editing studies, and opposed the clinical use of heritable gene editing in people.

But then, in June, another scientist with ambitions to produce gene-edited babies spoke up. Russian molecular biologist Denis Rebrikov told *Nature* that he was considering implanting gene-edited embryos into women (see page 372). Most recently, he expressed interest in repairing a mutation linked to deafness, and said that he had started experiments to investigate. But he also said that he will wait until Russian regulatory authorities grant permission before implanting gene-edited embryos.

While debate rages around genome editing in the clinic, researchers have continued to improve on the technology. In October, a team led by chemical biologist David Liu of the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, unveiled a method called prime editing. Early results suggest that this alternative tool could be more precise and accurate than standard CRISPR–Cas9 editing, which might ease some of the concerns about the safety of using gene editing in humans.

Indecisive leadership

In the United Kingdom, universities have stockpiled supplies as the country reached the brink of a no-deal Brexit this year, only for



Google's Sycamore quantum processor.

GOOGLE/REUTERS

its government to extend the deadline for leaving the European Union three times. Neither prime minister Theresa May nor her successor Boris Johnson managed to secure the backing of Parliament for a Brexit deal, and the ongoing uncertainty continues to worry scientists. To strengthen the country's science base, Johnson's government – which was re-elected on 12 December and now plans to exit the EU through the deal Johnson negotiated – promised to double government funding for research and development to £18 billion (US\$24 billion) a year by 2025, and to introduce a visa scheme that is more favourable to researchers.

In the United States, several science agencies started the year in suspended animation – caught up in a partial government shutdown that lasted a record-setting 35 days. NASA and the National Science Foundation (NSF) were among the agencies forced to halt most activities. Lawmakers did not resolve the impasse until late January.

Amid the chaos, the US Senate confirmed meteorologist Kelvin Droegemeier to serve as President Donald Trump's science adviser and lead the White House Office of Science and Technology Policy. Trump had gone nearly two years without a science adviser. Droegemeier quickly became a key player in the push to root out undue foreign influence in US science. Since 2018, the National Institutes of Health has investigated at least 180 scientists for failing to declare ties to foreign governments; many of the researchers are Chinese American, prompting fears that they were being unfairly targeted because of their ethnicity. Meanwhile, the Department of Energy and the NSF moved to bar their employees from participating in foreign talent-recruitment programmes.

Australia, too, has cracked down on foreign interference. In August, the government announced plans for an expert committee to



Health workers grappled with an Ebola outbreak in the Democratic Republic of the Congo.

respond to cyberattacks, intellectual-property theft and other strikes against universities by foreign governments or groups.

Elsewhere, scientists have found themselves caught up in civil unrest. In Hong Kong, violent clashes between police and protesters disrupted teaching and research on three university campuses (see page 383). And Chile had to pull out of hosting the United Nations COP25 climate summit because of safety concerns caused by massive protests against economic inequality in Santiago. The December talks were eventually held in Madrid.

A picture of health

An ongoing Ebola outbreak in the eastern Democratic Republic of the Congo (DRC) flared throughout the year, and has killed more than 2,200 people since it began in August 2018.

This is the second-worst Ebola epidemic yet in terms of deaths, and is the most complicated to address, owing to ongoing conflict in the region (see page 367). Ebola responders have been attacked by armed groups, and widespread mistrust of government officials and aid workers causes many residents to avoid treatment centres. In July, the World Health Organization declared the outbreak a “public health emergency of international concern” – its highest alert level.

Despite the chaos, researchers managed to conduct the first large, controlled trial of four experimental Ebola drugs. They found that two antibody-based therapies cured 90% of people who sought treatment in the early stages of the disease. And health workers have given more than 256,000 people in the eastern DRC a new Ebola vaccine manufactured by the pharmaceutical company Merck. In November, the vaccine became the first in the world to gain approval by a medicines agency.

In the United States, an outbreak of lung injuries in users of electronic cigarettes has killed more than 50 people and hospitalized more than 2,000, sending researchers and public-health officials scrambling to find the cause.

And in March, a person with HIV (whose identity hasn't been disclosed) was declared free of the virus after a stem-cell transplant swapped their white blood cells with HIV-resistant versions. They are only the second patient to have been successfully treated using this method after the ‘Berlin patient’, Timothy Ray Brown, was reported free of both HIV and leukaemia in 2009.

Written by Davide Castelvocchi, David Cyranoski, Elizabeth Gibney, Heidi Ledford, Amy Maxmen, Lauren Morello, Emma Stoye, Nidhi Subbaraman, Jeff Tollefson and Alexandra Witze.



UK Prime Minister Boris Johnson pledged to raise spending on research and development.

365 days the year in science

IMAGES OF THE YEAR

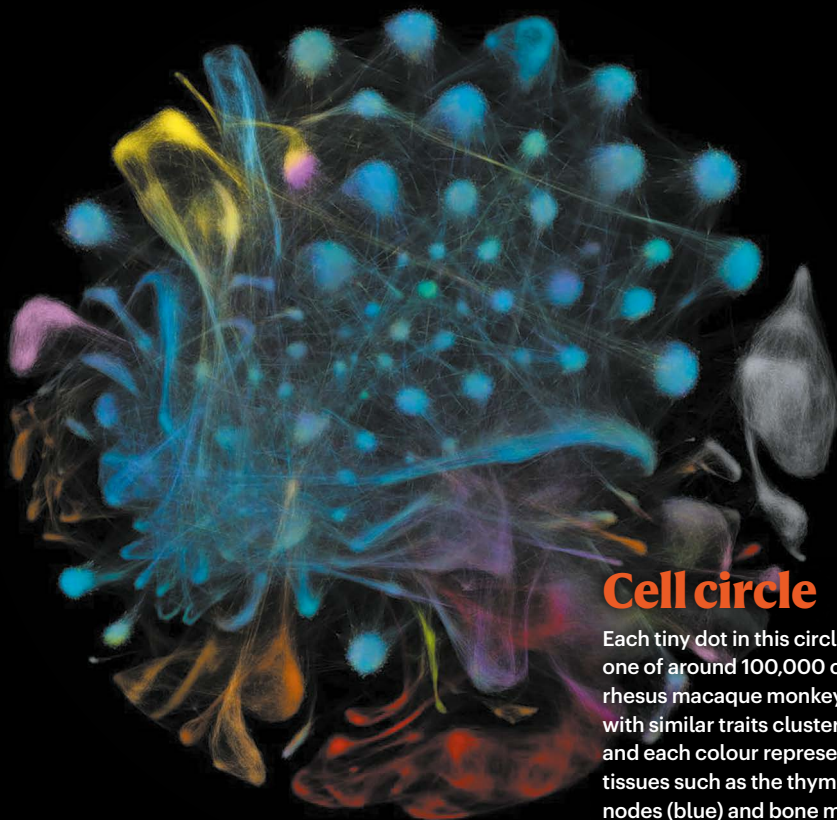
2019 will be remembered as the year humanity captured the first-ever image of a black hole. The year also brought fresh views of some of Earth's smallest living creatures and ominous signs of its changing climate. Here are the most striking shots from science and the natural world that caught the attention of *Nature's* news team.

Images selected by *Nature's* art editors
Text by Nisha Gaind and Ewen Callaway



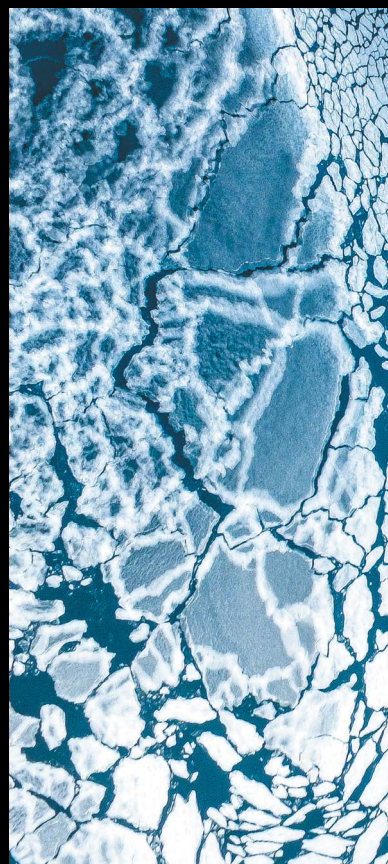
Go with the flow

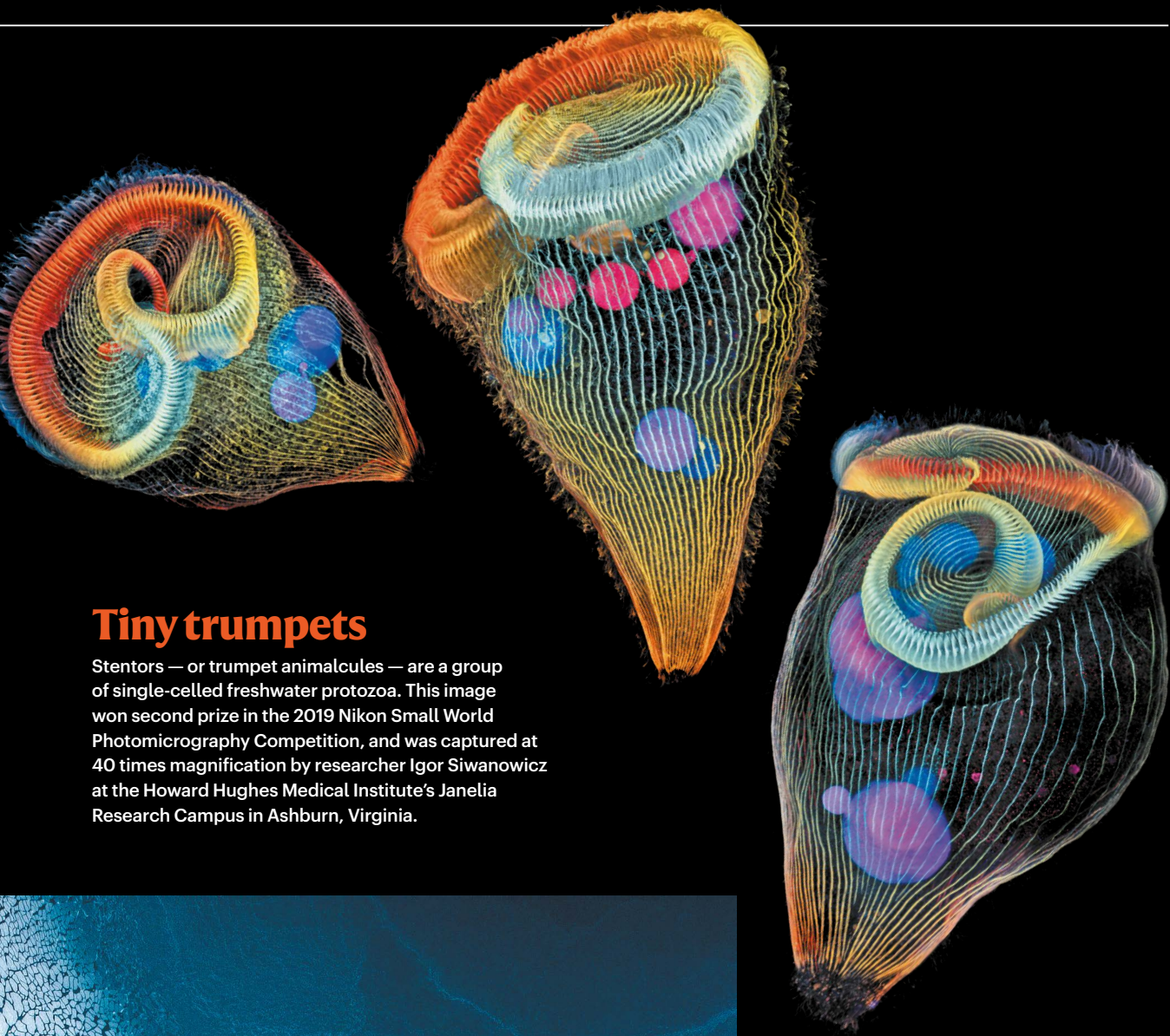
French researchers carved a labyrinth of microfluidic chambers in a silicon wafer to mimic blood flows in circulatory networks. Biophysicist Benoît Charlot at the University of Montpellier, France, captured this image using a scanning electron microscope.



Cell circle

Each tiny dot in this circle represents one of around 100,000 cells from rhesus macaque monkeys. Cells with similar traits cluster together, and each colour represents different tissues such as the thymus and lymph nodes (blue) and bone marrow (red).





Tiny trumpets

Stentors — or trumpet animalcules — are a group of single-celled freshwater protozoa. This image won second prize in the 2019 Nikon Small World Photomicrography Competition, and was captured at 40 times magnification by researcher Igor Siwanowicz at the Howard Hughes Medical Institute's Janelia Research Campus in Ashburn, Virginia.



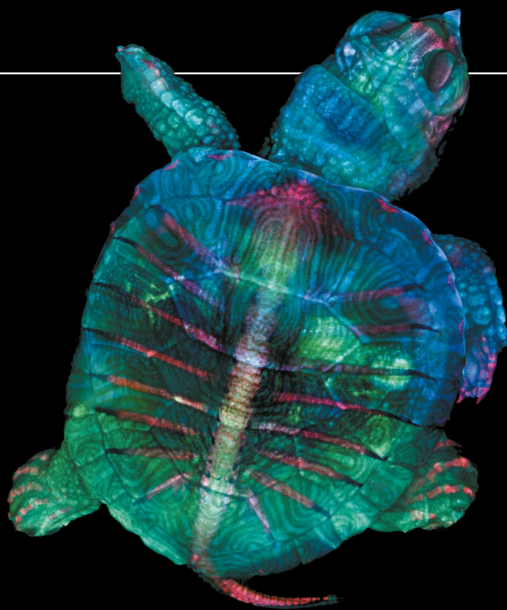
On thin ice

This aerial view of the sea ice in East Greenland was captured by photographer Florian Ledoux using a drone. Low levels of winter snow cover, heatwaves in the spring and a sunny summer all contributed to significant melting of Greenland's ice sheet in 2019.



Hole horizon

The Event Horizon Telescope collaboration unveiled this first direct image of a black hole and its event horizon in April. The team used eight radio observatories to capture the ring of light around the void, which is at the centre of the galaxy Messier 87.

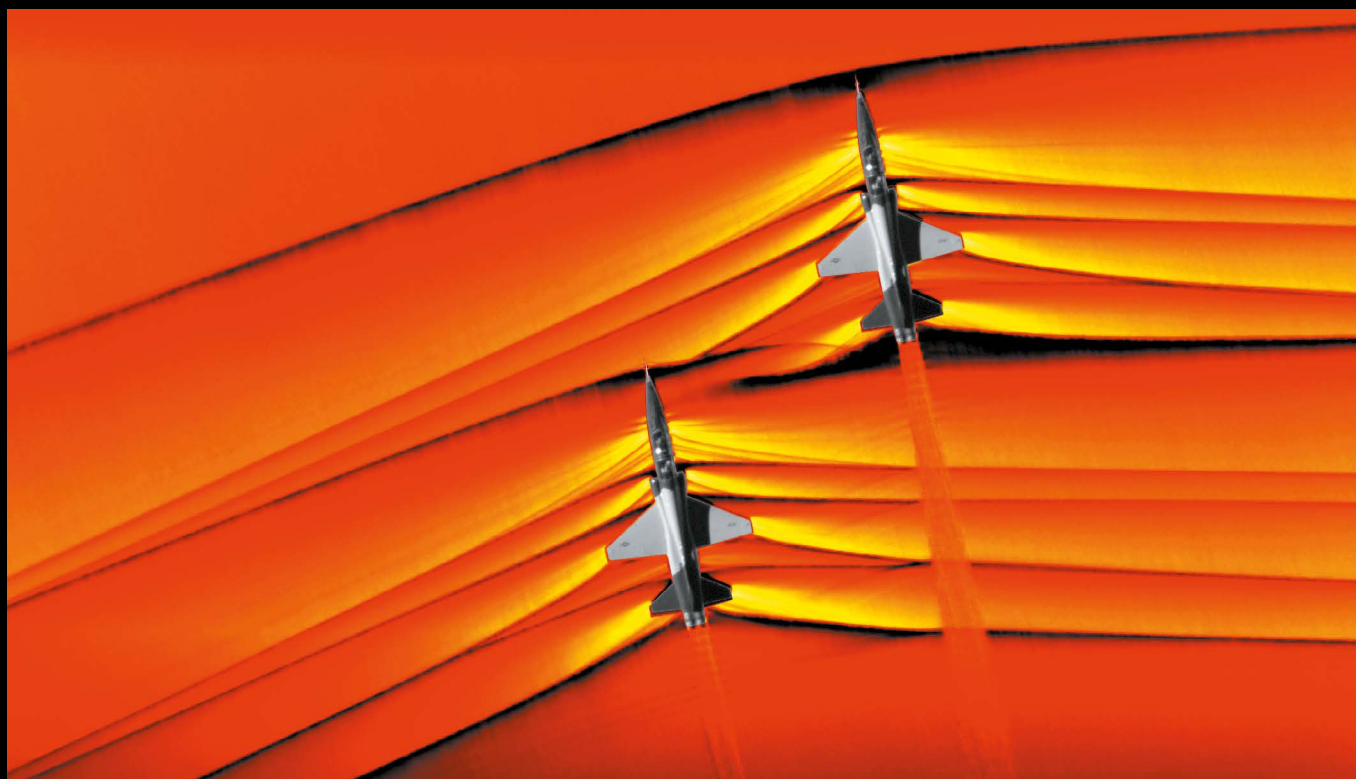


Out of its shell

This fluorescent visualization of a turtle embryo was the winner of the 2019 Nikon Small World Photomicrography Competition. Microscopists Teresa Zgoda and Teresa Kugler stitched together and stacked hundreds of stereomicroscope images of the roughly 2.5-centimetre-long embryo.

Friends reunited

Astronaut Christina Koch took this picture of the Soyuz spacecraft carrying her NASA colleague Jessica Meir as it approached the International Space Station (ISS). On 18 October, the pair performed history's first all-female spacewalk, to repair a faulty battery unit on the ISS.



Good vibrations

This false-colour image shows shockwaves emanating from supersonic US T-38 Talon aircraft, and was captured by NASA staff using an experimental technique from an aeroplane above. It shows the rapid changes in air pressure that cause people to hear sonic booms. The data will help aeronautical engineers to design quieter supersonic planes.

BLACK HOLE: ETH COLLABORATION, BLOOD FLOW MIMIC: BENOIT CHARLOT/CNRS/IES UNIV. MONTPELLIER/LABEX NUMEV, PROTOZOA TRUMPETS: IGOR SIWANOWICZ, CELL CIRCLE: CARLY ZIEGLER, ALEX SHALEK AND SHAINA CARROLL (MIT) AND LESLIE KEAN, VICTOR TRACHEV AND LUCREZIA COLONNA (DANA-FARBER CANCER INSTITUTE)/WELLCOME PHOTOGRAPHY PRIZE 2019, ICE SHEET: FLORIAN LEDOUX, TURTLE: TERESA ZGODA AND TERESA KUGLER, SOYUZ FROM ISS: CHRISTINA KOCH/NASA, SHOCKWAVE: NASA, ZIMBABWE ANTI-POACHING: BRENT STIRTON/GETTY, BLEACHED ANEMONE: MORGAN BENNETT-SMITH, WHISKY DROPLET: STUART J. WILLIAMS.



Brave one

Petronella Chigumbura is a member of the Akashinga, or 'brave ones', an all-female anti-poaching unit. They patrol Zimbabwe's Phundundu Wildlife Area in the Zambezi Valley, where elephant poaching is common.

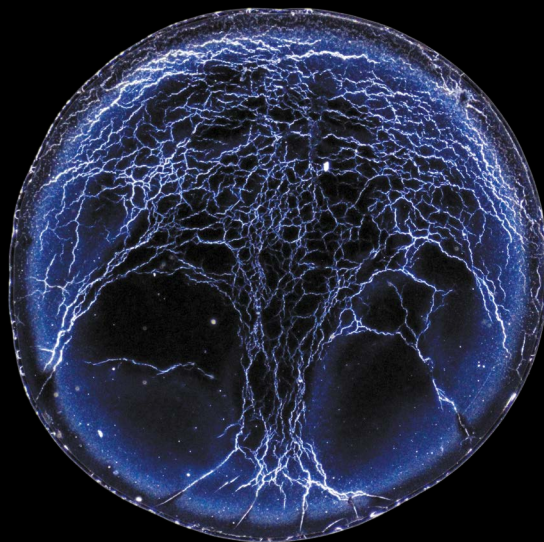


Anyone home?

A fish explores a bleached sea anemone in the Red Sea, off the coast of Saudi Arabia. Like corals, anemones form symbiotic relationships with algae that are disrupted when oceans get too warm, causing the anemone to expel the algae and become colourless.

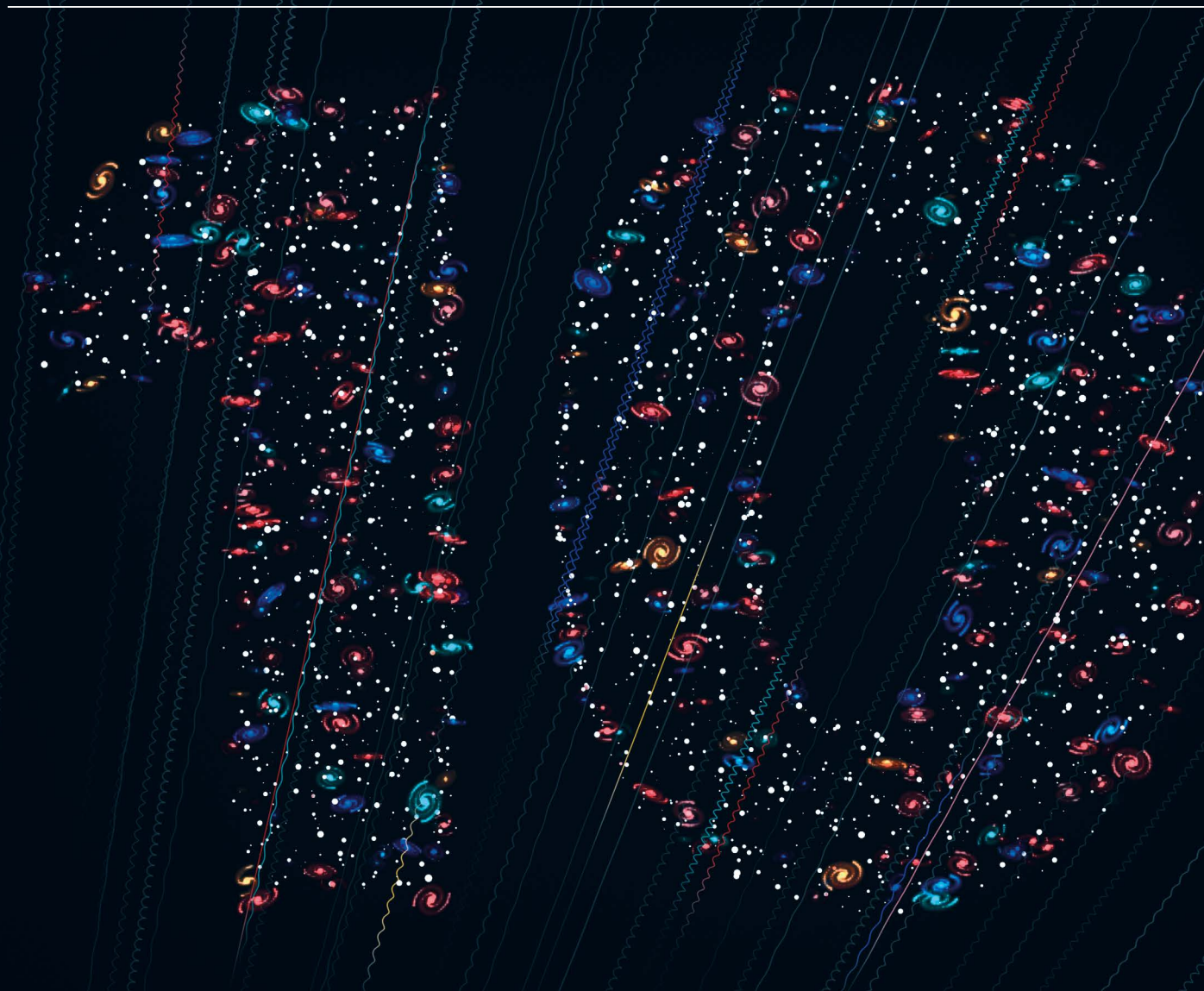
Wee dram

This web-like microstructure is made of fats left behind after researchers evaporated a 1-microlitre drop of diluted bourbon whisky. The fats dissolve in higher-strength spirits — but turn a drink cloudy when water is added.



NATURE'S 10

Ten people who mattered this year.



Ricardo Galvão / Victoria Kaspi / Nenad Sestan / Sandra Díaz /
Jean-Jacques Muyembe Tamfum / Yohannes Haile-Selassie /
Wendy Rogers / Hongkui Deng / John Martinis / Greta Thunberg

Nature's 10 is the journal's annual list of ten people who mattered in science this year. They might have achieved amazing discoveries, brought attention to crucial issues, or even gained notoriety for controversial actions. Although not an award or a ranking, *Nature's 10* highlights individuals who had a role in some of the year's most significant moments in science.

Ricardo Galvão

Science defender

As chaos spiked in the Amazon, the physicist became a national hero by challenging Brazil's government.

By Jeff Tollefson

Ricardo Galvão nearly passed out when he heard the news and realized he was being targeted by his own president.

On 19 July, Brazil's leader, Jair Bolsonaro, lashed out against a report on deforestation by Galvão's team at the National Institute for Space Research (INPE) in São Paulo. The group's analysis had incited the president's wrath because it found a sharp spike in forest clearing in the Amazon. The president accused the scientists of lying about the data and suggested that Galvão – as head of the institute – might be in cahoots with environmentalists. The 72-year-old fusion physicist was stunned by the accusation. "My wife had to bring me a glass of water," he says.

Rather than rush to react, Galvão gave himself 12 hours to craft a response. After a nearly sleepless night, he spoke out in defence of INPE scientists. He also accused the president of cowardice and called for a face-to-face meeting – acts that he knew would lead to him losing his job. What he didn't know was that he would become a hero of sorts, hailed by his scientific colleagues as well as by strangers on the streets. A woman even stopped him on the subway in São Paulo to thank him for standing up to Bolsonaro and helping her to understand why preserving the Amazon matters.

"He lost his job because he took a very clear and strong position in defence of science – and against authoritarianism," says Paulo Artaxo, an atmospheric physicist and Galvão's colleague at the University of São Paulo. Artaxo sees worrisome parallels between Bolsonaro's government and the dictatorship that ruled Brazil between 1964 and 1985, including a tendency to attack any evidence that doesn't support its political goals. "We

need people like Galvão to stand up."

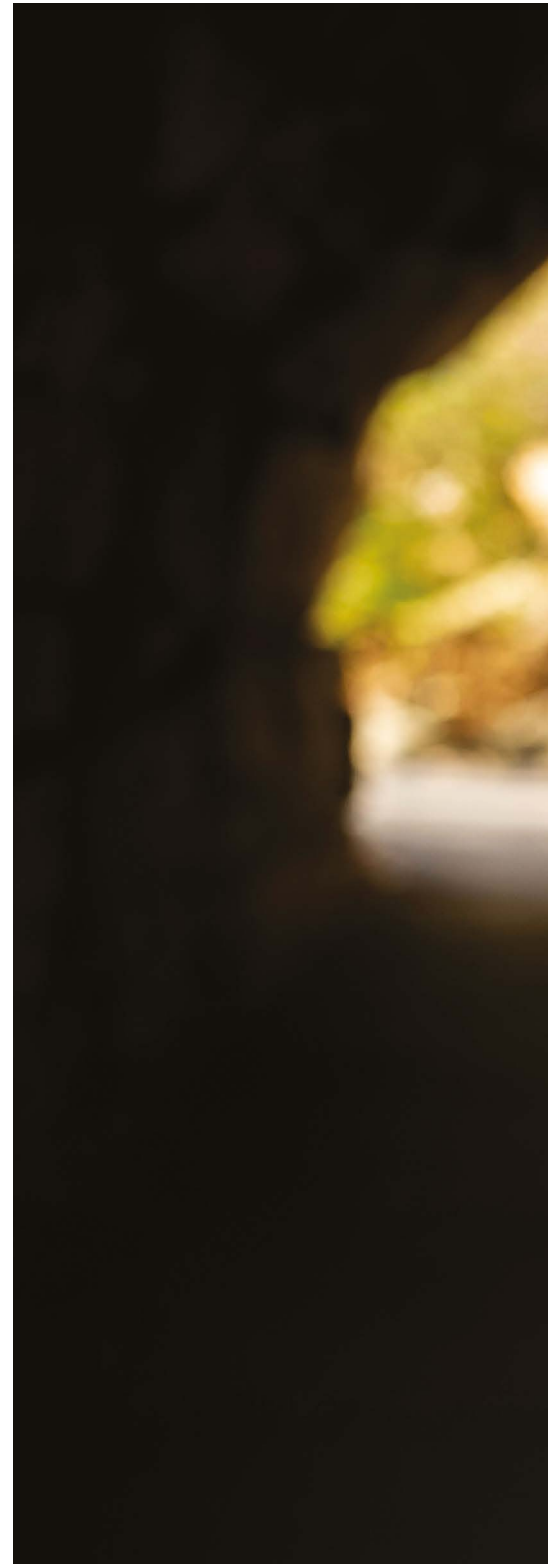
It wasn't Galvão's first run-in with the Bolsonaro administration. Officials had repeatedly questioned the accuracy of INPE's deforestation alerts, which use detailed analysis of satellite imagery.

This time, however, the president was attacking the integrity of scientists and one of Brazil's top scientific institutions. As expected, Galvão was dismissed two weeks after he defended INPE, just as the burning season kicked off in the Amazon. Farmers light fires as the last step in clearing the land for agriculture.

Brazil's reputation as an environmental leader has been deteriorating in recent years. The country had managed to curb deforestation by more than 80% between 2004 and 2012, but the aggressive environmental enforcement ended up sparking a political backlash and a rise in deforestation.

INPE's latest numbers, released on 18 November, show that an estimated 9,762 kilometres of land – an area larger than Puerto Rico – was cleared between August 2018 and July 2019. That is an increase of 30% over the previous year, and more than twice the area cleared in 2012. Scientists and conservationists charge that Bolsonaro's anti-environmental rhetoric has sent a signal to ranchers, farmers and land-grabbers that they can once again clear forest in the Amazon with impunity.

Galvão has since returned to his previous position at the University of São Paulo. He doesn't enjoy the limelight and was preparing to stop giving interviews and focus on his fusion research. After receiving messages from fellow scientists thanking him for speaking out, however, he realized that he has a responsibility to continue to advocate on behalf of science – and scientists – in the face of political pressure. "I'm just a humble old man who works in physics," Galvão says. "But I decided to go on for this reason."



**“He took a very clear
and strong position
in defence of science
— and against
authoritarianism.”**



MICHAEL RUBIN FOR NATURE

Victoria Kaspi

Sky sleuth

An astrophysicist chased mysterious fast radio bursts with an innovative radio telescope.

By Alexandra Witze

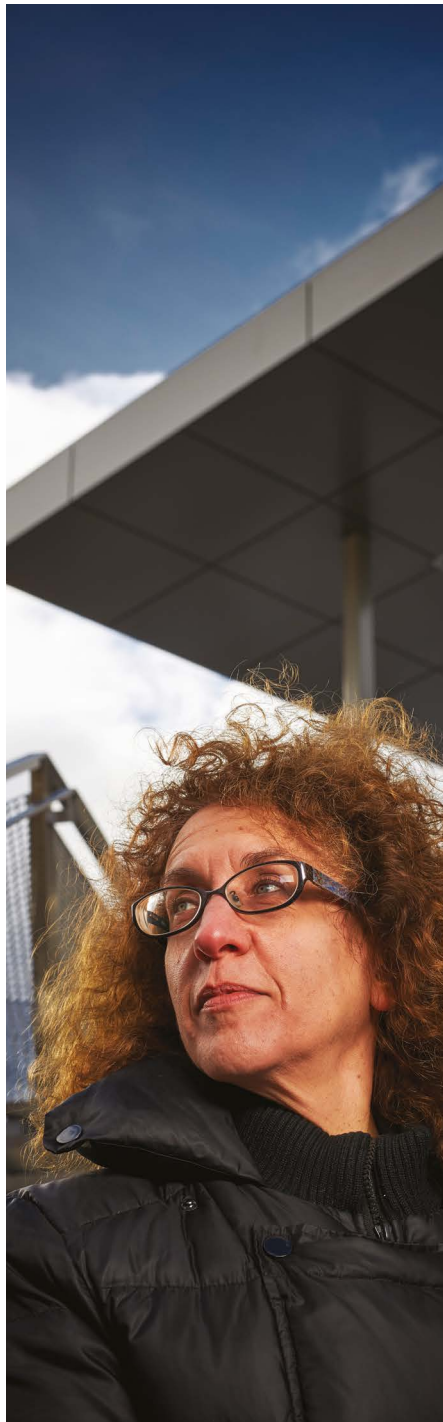
In the past quarter of a century, Victoria Kaspi has used many of the world's top telescopes to make fundamental astronomy discoveries. But in 2017, she found herself helping to build one, screwing in hundreds of cables to connect the Canadian Hydrogen Intensity Mapping Experiment (CHIME) to powerful computers.

This year, the efforts of Kaspi and dozens of other astronomers paid off. CHIME became the world's best hunter of fast radio bursts (FRBs) – mysterious flashes of radio energy that frequently pop off across the sky. CHIME, which is located in southern British Columbia, has spotted hundreds of bursts, many more than any other telescope. With it, astronomers hope to solve the puzzle of the signals' origin.

Kaspi, an astrophysicist at McGill University in Montreal, Canada, had a major role in giving CHIME its powerful FRB-detection capabilities. The telescope was originally designed to map hydrogen emission from distant galaxies, to answer questions about the early Universe. But as the project was coming together in the early 2010s, so, too, was the burgeoning field of FRBs, the first of which was discovered in 2007. In 2013, astronomers reported four more examples, confirming that the flashes were a real phenomenon that needed explaining.

"That moment to me was a watershed," says Kaspi. She had spent much of her career studying ultra-dense stellar remains known as neutron stars. But, suddenly, a new astrophysical mystery was emerging. Kaspi had been thinking about how CHIME could study fast-rotating neutron stars and realized the telescope's sensitivity and large field-of-view could be ideal for bagging FRBs – but only if it were upgraded. She telephoned Ingrid Stairs, an astronomer at the University of British Columbia in Vancouver, to chew over the idea. "And within a few months, she was leading this big proposal," says Stairs.

Kaspi worked with the cosmologists who



"At this point, it's like drinking from a fire hose, we have so much data."

dreamt up the telescope to request more money from its main funder, the Canada Foundation for Innovation in Ottawa, to hunt for FRBs. They wanted to add another instrument and enough computing power to enable the telescope to churn through data gathered 1,000 times per second at 16,000 different frequencies. "We all knew this was very risky," she says. "The telescope hadn't been built yet, and here we were proposing to add something onto something that didn't exist."

But Kaspi, who became principal investigator of the FRB part of CHIME, pulled it off. Her scientific chops helped win the funding; her personal connections enabled her to build a large and diverse team; and her political skills were essential in bringing together the original cosmologists and the new FRB hunters, says Matthew Bailes, an astronomer at Swinburne University of Technology in Melbourne, Australia.

Along the way, Kaspi has worked to develop the next generation of scientists, mindful of how challenging it can be to enter physics, especially for women. She won the nation's highest science prize, the Gerhard Herzberg Canada Gold Medal for Science and Engineering, in 2016 and used the Can\$1-million (US\$760,000) prize to hire students and postdocs for CHIME.

This year, she helped to land a US\$2.4-million grant from the Gordon and Betty Moore Foundation to explore building 'outrigger' telescopes. The outriggers would be sited some 1,000 kilometres away from CHIME and help in pinpointing FRBs. That would keep this inventive Canadian telescope at the forefront of astronomy.

"At this point, it's like drinking from a fire hose, we have so much data," Kaspi says. "Honestly, I'm blown away."

CHRISTIAN FLEURY FOR NATURE

Nenad Sestan

Brain rebooter

“I realized it was something beyond our expectations.”

A neuroscientist revived disembodied pig brains and challenged definitions of life and death.

By Sara Reardon

Nenad Sestan was working in his office one afternoon in 2016, when he heard two of his lab members in a room across the hall giggling with excitement over a microscope. “I knew something was happening,” he says. “I realized it was something beyond our expectations.”

The researchers, at Yale School of Medicine in New Haven, Connecticut, had found electrical activity in brains taken from dead pigs. The team had painstakingly removed the organs shortly after death and infused them with oxygen and an ice-cold preservative, and in doing so, brought the brains at least partially back to life. With that shocking result, Sestan realized that what had started as a side project to find ways to better preserve brain tissue for research had morphed into a discovery that could redefine our understanding of life and death.

The excitement soon turned to concern, when the researchers thought they saw widespread, coordinated electrical activity – the type that can indicate consciousness. Sestan brought in a neurologist, who determined that the readout was actually an error, but the possibility had spooked them.

Sestan kept his cool and immediately did two things: he shut down the experiment



and contacted the US National Institutes of Health (NIH), which funds his research, as well as a Yale bioethicist. Over the next few months, experts pored over the potential ethical implications, such as whether the brains could become conscious and whether physicians needed to reconsider the definition of brain death.

Sestan had anticipated the ethical questions and adopted some safeguards. Before starting the experiments, the group had decided to anaesthetize the brains with blocker drugs to prevent neurons from firing in unison – a prerequisite for consciousness.

Overall, the feat met with more excitement than concern. Sestan's results suggested that oxygen deprivation, which can happen during a stroke or severe injury, was not as damaging to brain cells as previously thought. “It's very important: something we overlooked, because nobody really thought that this was possible,” says Anna Devor, a biomedical engineer at Boston University in Massachusetts.

Once they were confident that the experiment was ethically sound, the researchers resumed their experiments. They submitted the work to *Nature*. But before the paper could be published, Sestan presented data at a public NIH neuroethics meeting and – despite his protests – the

story appeared in the press.

Sestan admits he was amused by some of the sensationalist headlines, dubbing his project ‘Frankenswine’ and ‘Aporkalypse’. But he was stung by suggestions that the researchers were engineering immortality, or maintaining a room full of living brains in jars. Neither he nor his team wanted to discuss the results until the paper was out, but as their inboxes filled with concerns and rants from animal-rights activists and futurists, Sestan became depressed. “We were really very worried,” he says. He felt that all they could do, however, was to hold off on correcting public misunderstandings until the peer-review process had run its course.

Since the paper was published in April (Z. Vrselja *Nature* **568**, 336–343; 2019), the team has been so busy fielding enquiries from the media and scientists that it hasn't performed any further experiments. Sestan wants to focus on his original questions and explore, for instance, how long the brains can be maintained for, and whether the technology can preserve other organs for transplantation.

From now on, this strand of his research will be decided by committee. “We want to get outside opinion before we do anything,” he says. “When you explore uncharted territory, you have to be very, very thoughtful.”

Sandra Díaz

Biodiversity guardian



An ecologist and her colleagues assess Earth's ecosystems and call for drastic action.

By Ehsan Masood

On 4 May, Sandra Díaz and 144 other researchers had a stark message for the world. They had just finished the most exhaustive study ever of the world's biodiversity, and the news was worse than most researchers had imagined: one million species are heading for extinction because of human activities, and it will take drastic action to stop that. "The rate at which species are going extinct is at least tens to hundreds of times faster than it has been on average over the past ten million years," Díaz says. "Our safety net is stretched almost to breaking point."

Those alarming findings came from the Intergovernmental Science–Policy Platform on Biodiversity and Ecosystem Services

(IPBES). Díaz, an ecologist at Argentina's National University of Córdoba, is one of the panel's three co-chairs. For most of the previous three years, she and her colleagues – anthropologist Eduardo Brondízio at Indiana University Bloomington and ecologist Josef Settele at the Helmholtz Centre for Environmental Research in Halle, Germany – coordinated the work of experts from 51 countries, meeting in physical

"We cannot live a fulfilling life, a life as we know it, without nature."

workshops and in virtual working groups, poring over 15,000 sources of information.

Their final report, which runs to 1,500 pages, says that nations will fail to meet most global targets in biodiversity and sustainable development unless they make massive changes, such as abandoning the idea that economies must grow constantly.

"We cannot live a fulfilling life, a life as we know it, without nature," Díaz says. And if economies continue to run in such a destructive way, a "new economic model is needed for nature and people", she says.

It's a blunt and, in some ways, radical message. But Díaz does not shy away from speaking out on important issues in science and policy. She challenged, for instance, what was once one of the central tenets in twentieth-century ecology: the idea that ecosystems and their benefits to humans – such as food, or climate regulation – depend heavily on having large numbers of species. Shahid Naeem, a researcher at Columbia University in New York City who studies the impacts of biodiversity loss, says that Díaz has led the charge in highlighting the value of what plants actually do – known as their functional traits.

This insight and others came to Díaz through years spent trudging through the fields of Africa, Asia, Europe and Latin America, collecting leaves, measuring their toughness, and assessing soil properties. It's a habit she developed growing up in central Argentina, when she would explore the Pampas grasslands while others took their afternoon rests. "I would escape the siesta to see plants and animals," she says. "Ever since I was an undergraduate, I knew I wanted to be a researcher."

Díaz has a second career now, beyond conservation science – influencing policy through her work with IPBES. She takes heart in how the panel's report is being adopted by many social and environmental movements, including Extinction Rebellion, that are pushing for stronger and more urgent action on the environment.

"We have been amazed at its reach," she says. "The report has come at the right moment." And despite its dark prognosis, Díaz refuses to be pessimistic about humankind's capacity to turn things around. "I have to be optimistic," she says, "because there is no Plan B".

Jean-Jacques Muyembe Tamfum

Ebola fighter

The co-discoverer of Ebola faces his tenth battle with the virus in the DRC – his toughest yet.



By Amy Maxmen

In 1976, Jean-Jacques Muyembe Tamfum travelled deep into the tropical forests of what is now the Democratic Republic of the Congo (DRC) to investigate an outbreak of an unidentified ailment that was swiftly killing people.

The young researcher realized that something was odd when he drew blood samples from those who were sick and the needle pricks wouldn't clot. Blood spilt over his bare hands. Nurses he worked beside were dying, and Muyembe began to worry. "I started taking my temperature every morning and every night," he says. Miraculously, he never fell ill from the virus, later named Ebola.

Now, 43 years after discovering the disease, Muyembe is leading the DRC's response to the most volatile outbreak of Ebola yet. Since August 2018, the epidemic has killed more than 2,200 people in the northeast of the country, a region already pummeled by a quarter of a century of conflict and political instability.

Muyembe – who took the helm of the response in July – brings deep experience to the effort, along with a dedication to

cutting-edge science. Beginning in 1995, he developed the key public-health measures still used to contain the virus. During a large outbreak in the DRC city of Kikwit, he realized that the most vital step was to converse with communities so that they trusted him and understood how to protect themselves. He found ways to bury the dead respectfully while minimizing the risk of infection. And he began investigations that would lead to the roll-out of effective Ebola drugs and vaccines. During this outbreak, he took blood from Ebola survivors and infused it into eight people who had been infected, in the hope that antibodies would quash the virus. Seven of the recipients survived.

Last month, a 680-person, controlled clinical trial led by his team showed a 90% survival rate for those treated with antibody-based drugs shortly after infection. One of the drugs, mAb114, is derived from an antibody from the blood of a survivor whom Muyembe recruited during the Kikwit outbreak. Nancy Sullivan, an immunologist at the US National Institutes of Health in Bethesda, Maryland, attributes the success to Muyembe's doggedness at the time. "His contribution was pivotal to show that you can do a trial in a chaotic outbreak," she says.

"He was there at the start and he is still there because he is so persistent."

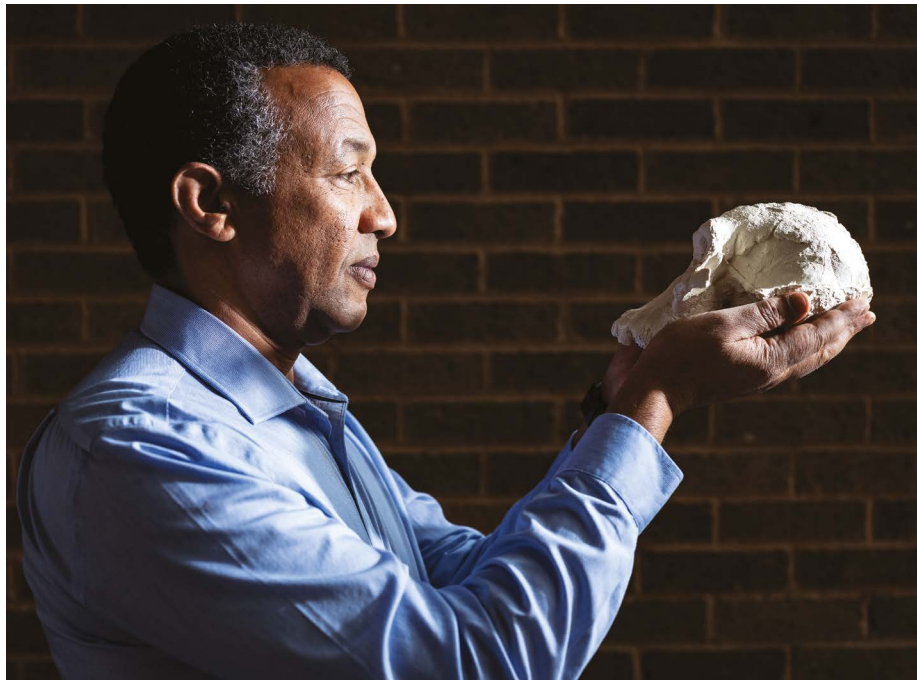
In recent weeks, the number of new Ebola cases being recorded has dwindled – progress that Muyembe's colleagues attribute in part to his leadership. Yet, in a rushed call with *Nature* in late November, the Ebola veteran was worried. Violence had broken out in Beni, a city hit hard by the epidemic, and Ebola responders were on lockdown. But backing down has never been an option for Muyembe. David Heymann, an epidemiologist at the London School of Hygiene and Tropical Medicine, says: "He was there at the start and he is still there because he is so persistent."

After this outbreak eventually ends, Muyembe is determined to unravel one last puzzle before retirement. His team has been collecting animals from regions where the virus has spilt into people, in the hope of tracking how the disease moves between species. "I want to find the vector," he says.

Yohannes Haile-Selassie

Origin seeker

A palaeontologist shook up the human family tree with the discovery of a remarkably preserved 3.8-million-year-old skull.



By Nisha Gaiind

It was a pale, circular shape on the ground, about the width of a grapefruit, that caught the attention of Yohannes Haile-Selassie when he was investigating a site in the northern Ethiopian desert in February 2016. The object was jutting out of the parched earth just 3 metres away from a jawbone found by a goat herder a few hours earlier. “Before I picked it up, I said, ‘Oh my goodness, this is something.’”

The fossils together formed a remarkably complete early hominin skull, which Haile-Selassie’s team dated to 3.8 million years old. It belongs to a species called *Australopithecus anamensis* – the oldest and most elusive known human relative. That afternoon, the team celebrated their rare find with cold Cokes, beers and dancing.

The skull, known as ‘MRD’ and revealed to the world in August (Y. Haile-Selassie *et al. Nature* 573, 214–219; 2019), gave researchers their first look at the face of

“Before I picked it up, I said, ‘Oh my goodness, this is something.’”

this enigmatic ancient relative, which was previously known from just a few bone fragments. Palaeoanthropologists are impressed by the specimen, and some say it is rivalled only by Lucy, the 3.2-million-year-old skeleton fossil of the closely related species *Australopithecus afarensis*. “That’s a big thing to hear from our colleagues,” says Haile-Selassie, a palaeoanthropologist at the Cleveland Museum of Natural History in Ohio.

Haile-Selassie is considered one of the field’s most talented fossil finders. Many treasures have surfaced from his project in Woranso-Mille, a region scattered with hominin fossils from the Pliocene, a key

period in the evolution of the genus *Homo* and its close relative *Australopithecus* between 5.3 million and 2.6 million years ago. He is also one of a crop of Ethiopian palaeoanthropologists who lead major scientific projects in their homeland – a big shift from a generation ago, when foreigners oversaw most of the research in this fossil-rich nation.

When Haile-Selassie was doing his PhD in the mid-1990s, his potential was clear and he had a knack for both laboratory and fieldwork, says Tim White, a palaeoanthropologist at the University of California, Berkeley, and Haile-Selassie’s PhD adviser. Fieldwork in remote areas is extremely difficult, says White, but Haile-Selassie has nailed it: organizing people, equipment, vehicles and permits, and all using multiple languages. “It’s not luck that Yohannes put these people in Ethiopia in this place, with all the right specialists to work on a problem.”

MRD is important in part because it is from a time period that was literally empty in terms of the fossil record, says Haile-Selassie. And it shook up the oldest branches of the hominin evolutionary tree. Researchers previously thought that Lucy’s species had evolved from *A. anamensis* – in a ‘classic’ case of direct evolution from one species to another, says Fred Spoor, a palaeontologist at the Natural History Museum in London. But Haile-Selassie and his colleagues argued that the skull’s features, together with the reanalysis of some existing fossils that it allowed, suggest that early hominin evolution was much messier, and that *A. anamensis* and *A. afarensis* overlapped for at least 100,000 years. It’s also vanishingly rare to find such an intact specimen. “The MRD find is an iconic cranium,” says White.

Not everyone agrees with the evolutionary shake-up that Haile-Selassie’s group has proposed. The team is still studying the cranium for more clues about its place in prehistory, and Haile-Selassie hopes to revisit the discovery site to enrich the picture. “Hopefully the rest of the skeleton might be there, who knows,” he says.

Of all his discoveries, MRD is number one, says Haile-Selassie. As a daily reminder, the moniker now features on his car’s number plate.

Wendy Rogers

Transplant ethicist

An academic revealed ethical failures in China's studies on organ transplants.



By David Cyranoski

For two decades, controversy has swirled around the origin of some livers, hearts and kidneys used for organ transplants in China. First, the government denied that organs had been taken from prisoners; then, it admitted it. It now says the practice has been banned since 2015, and that organs all come from volunteers. But researchers have questioned that, too.

Wendy Rogers, a bioethicist at Macquarie University in Sydney, Australia, found a new way to prise open the issue: examining research publications by Chinese transplant doctors. Her team's investigation, published in February (W. Rogers *et al. BMJ Open* 9, e024473; 2019), triggered more than two dozen retractions of reports of transplants, after doctors couldn't prove that donors gave consent. "If you think about what's really happening, it's unbearable," Rogers says.

The retractions help to place the practice among the world's major bioethical scandals, says Yves Moreau, a computational biologist at the Catholic University of Leuven in Belgium — and show how seriously scientists and publishers should take research ethics.

Rogers's shift from academic to activist

started at a 2015 conference that screened a documentary, *Hard to Believe*, discussing forced organ donations from political prisoners. Rogers had studied Australia's transplant system, and was shocked by what was going on in China. In 2016, she became the unpaid chair of the international advisory committee of the International Coalition to End Transplant Abuse in China (ETAC), a non-profit advocacy group in Sydney. Following an anonymous lead, she investigated a 2016 paper in *Liver International*, in which she found the documentation of donors lacking; the paper was retracted in 2017.

Rogers knew there must be many more problematic papers. She worked over nights and weekends with a team of researchers and volunteers to sift through thousands of papers. They found more than 400 that Rogers's team concluded had probably used organs from prisoners and didn't make their source clear. Those papers, published between 2001 and 2017, reported more than 85,000 transplants. The team spotlighted 17 journals that had published 5 or more papers. Two reacted. *PLoS ONE* retracted 19 of the 21 papers on Rogers's list; an investigation into the other two is ongoing. *Transplantation* retracted seven: five of six on Rogers's list and

two it identified on its own. The retraction notices say that the authors didn't respond or couldn't give satisfactory explanations.

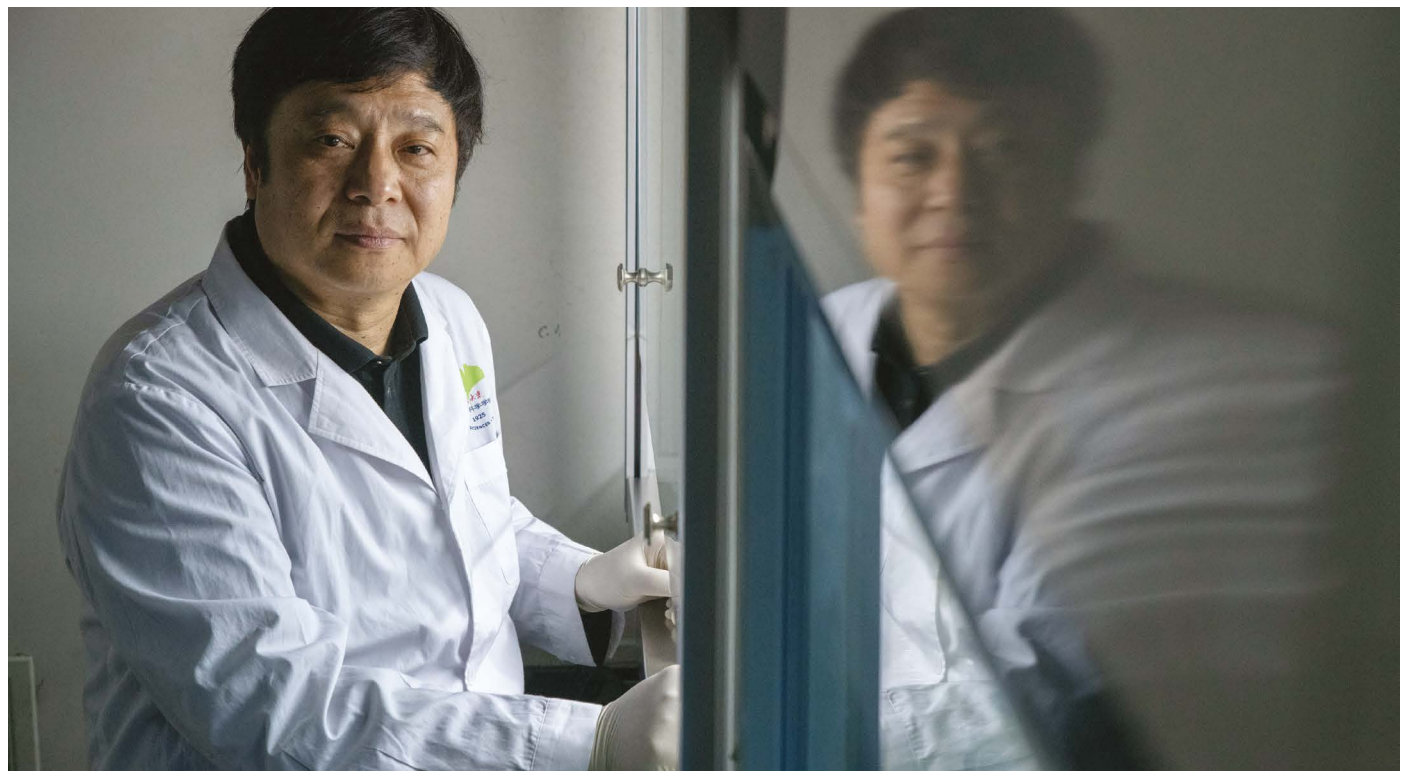
Joerg Heber, editor-in-chief of *PLoS ONE*, says he is grateful to Rogers's team. His journal has now strengthened its reporting requirements for transplantation papers. For the journals that haven't responded, Rogers says, "I really urge them to take this seriously."

ETAC approached Geoffrey Nice, a lawyer with experience prosecuting war criminals in The Hague, the Netherlands, to write about what was happening. Nice suggested an international panel, which he chaired and to which Rogers gave evidence. The panel also considered a paper published this year that questioned data from China's organ donation programme (M. P. Robertson *et al. BMC Med. Ethics* 20, 79; 2019).

In June, the panel concluded that people imprisoned for their religious or political views had been killed for their organs in China, and that the practice probably continues. That report and Rogers's work were both met with silence from China. Rogers is not optimistic that China will ever be fully transparent about its transplants, but the scrutiny might stop any forced harvesting of organs, she says.

Hongkui Deng

CRISPR translator



A Chinese scientist shows that CRISPR gene editing can be used safely in adults with HIV.

By David Cyranoski

The CRISPR–Cas9 gene-editing system was developed less than a decade ago, and it is already showing up in the clinic. This year marked what many consider the first published report of its use in a person. The study came from the laboratory of Hongkui Deng at Peking University in Beijing, and showed how CRISPR gene editing can create a potentially limitless supply of immune cells that are impervious to infection by HIV.

The approach was designed to recapitulate the success with the ‘Berlin patient’, Timothy Ray Brown. In 2008, Brown became the first person known to be cleared of the virus, thanks to a bone-marrow transplant he received as part of a treatment for leukaemia. His doctors had intentionally sought a donor with a genetic mutation that disables CCR5, a protein that HIV uses to infect immune cells. They wiped out Brown’s immune system,

then replenished it with the donor cells, and the virus disappeared.

But the protective mutation found in Brown’s bone-marrow donor is rare, and it is practically non-existent in China. So Deng, who was part of one of the teams that discovered CCR5’s importance in HIV in the 1990s, decided to try editing the gene instead. He took immunologically matched blood-forming stem cells from the bone marrow of a donor, edited them with CRISPR–Cas9 and then transplanted them into a person with leukaemia and HIV. “We are hoping for an exact mimicry of the Berlin patient,” he says.

But for safety’s sake, and because the type of cell used is difficult to edit, Deng used a mixture of cells for the transplant; only about 18% had been modified. The patient’s HIV infection remained (L. Xu *et al.* *N. Engl. J. Med.* **381**, 1240–1247; 2019). Deng says the work showed how CRISPR–Cas9 edited cells could be part of a bone-marrow transplant

and not cause adverse events. Some edited cells remain in the patient’s blood today, almost two years later. “We were most interested in persistence,” he says. “That is the major result of the paper.”

But according to Fyodor Urnov, a biologist at the University of California, Berkeley, who also is trying to push CRISPR into the clinic, the failure to successfully treat HIV was indicative of a rush to translate the technology. Previous experiments had shown that the clinical benefits achieved through other gene-editing methods, such as zinc-finger nucleases, depend on the efficiency of the editing process. The treatment couldn’t possibly work with such a low percentage of edited cells, Urnov says, and Deng and his team should have known that. “Their work will stand as an example of how not to do this,” he says.

Deng defends the experiment, and hopes in the short term to transplant a higher proportion of gene-edited cells. He also wants to develop methods for reprogramming cells into pluripotent stem cells, which are easier to edit, and then convert them into blood-forming stem cells for transplant. “That would be perfect,” says Deng.

John Martinis

Quantum builder

“It was clear to me that this was a great idea and that it would be wonderful to work on it.”

A physicist led Google’s first demonstration of a quantum computer that could outperform conventional machines.

By Elizabeth Gibney

When John Martinis was a graduate student in the mid-1980s, he went to a lecture that set the course of his scientific life. The famous physicist Richard Feynman discussed the idea of using particles’ quantum characteristics to make computers that could do things that are impossible on conventional machines. “It was clear to me that this was a great idea and that it would be wonderful to work on it,” Martinis says.

In October, Martinis took a big step towards Feynman’s dream. He led the work of a group of researchers at Google who announced that they had demonstrated a first: a quantum computer that could carry out a calculation faster than the best conventional computer. “Doing this experiment was the culmination of my career,” says Martinis.

The physicist, who works both at Google and at the University of California, Santa Barbara, has spent 17 years honing the hardware that underpins the firm’s quantum computer, named Sycamore. At its heart are tiny superconducting loops known as qubits, quantum systems that seem to exist in multiple states until they are observed. Physicists have long theorized that harnessing interactions between qubits could enable computers to excel at certain calculations, such as probing otherwise unsearchable databases and cracking conventional encryption.

A team of more than 70 scientists and engineers showed that, for a specific challenge – calculating the spread of outputs from a kind of quantum random-number generator – Sycamore could do in 200 seconds what they estimated would take the best supercomputer 10,000 years (although others argued that it would need only days).

The feat relied on improved hardware that lowered error rates and connected qubits in new ways. Some physicists debated

the significance of the landmark, and the task has limited practical application. But Martinis says the experiment’s importance lies in demonstrating something fundamental: that physicists’ understanding of quantum interactions – learnt on small quantum systems – remains true at larger scales and complexity. “That’s really good news,” he says.

Hartmut Neven, who leads Google’s Quantum Artificial Intelligence laboratory, says that Martinis used to be a mountain

climber, and that he applies that same careful, deliberate approach to building hardware, in which every sequence of moves must be thought out in advance. “John’s idea of a relaxing Sunday is to go into the lab and solder something together,” he says. “Life and work aren’t really separated.”

Martinis has many more ideas he’s hoping to pursue. His future priorities include making better quantum chips – including mastering methods to correct for errors caused by noise – and opening up Sycamore for use by outside researchers on a cloud system, to see whether there are useful algorithms that it could run. One idea is a method to verify that supposedly random numbers are truly random. “Physicists like me don’t retire,” he says with a smile. “We have lots of things to do.”



ONES TO WATCH 2020

António Guterres

Secretary-general, United Nations

This Portuguese diplomat has pushed for aggressive action on global warming, and that advocacy could prove crucial as nations meet in 2020 to update their pledges under the 2015 Paris climate agreement.

Denis Rebrikov

Kulakov National Medical Research Center for Obstetrics, Gynecology and Perinatology, Moscow

There has been an international outcry over this biologist's plans to start producing gene-edited babies, pending approval from the Russian government.

Geng Meiyu

Shanghai Institute of Materia Medica, China

This researcher and her team discovered a compound that has been approved in China to treat Alzheimer's disease, but sceptics have questioned its efficacy and many await more definitive trials.

Mariya Gabriel

European Commissioner for Innovation, Research, Culture, Education and Youth

The European Union's next spending plan for research and innovation, known as Horizon Europe, will take shape under the leadership of this Bulgarian political scientist.

Markus Rex

Alfred Wegener Institute, Germany

This atmospheric scientist is leading the €140-million (US\$155-million) MOSAiC expedition, in which a German ship frozen into Arctic ice for a year will collect data about polar conditions.

Greta Thunberg Climate catalyst

A Swedish teenager brought climate science to the fore as she channelled her generation's rage.



By Quirin Schiermeier

At a US congressional hearing on climate change in September, Greta Thunberg slid a slim bundle of papers across the table towards lawmakers. It was a special report from the Intergovernmental Panel on Climate Change, predicting dire consequences as the world warms. "I don't want you to listen to me, I want you to listen to the scientists," she told the legislators. "I want you to unite behind the science and I want you to take real action."

Scientists have spent decades warning about climate change, but they couldn't galvanize global attention the way that Thunberg did this year. The Swedish 16-year-old has outshone them – and many are cheering her along.

"Some may wonder why a teenage girl should get more credit and attention for publicly lamenting a well-known dilemma than most climate researchers get for years of hard work and effort," says Sonia Seneviratne, a climate scientist at the Swiss Federal Institute of Technology in Zurich. But Thunberg is candid and her outrage

unvarnished, and that is powerful, says Seneviratne. "As scientists, we normally don't dare to express the truth in such heartfelt simplicity."

Many researchers hail Thunberg in particular for focusing attention on climate change and its catastrophic impacts. What she has achieved should motivate climate researchers to carry on with their science despite slow political action, says Seneviratne.

"Greta has inspired scientists along with activists and policymakers," says Angela Ledford Anderson, director of the Climate and Energy programme at the Union of Concerned Scientists in Washington DC. In July, German Chancellor Angela Merkel announced sweeping measures to reduce carbon emissions, and acknowledged that the protests Thunberg ignited "drove us to act".

But perhaps Thunberg's biggest influence will be on the next generation of scientists, Anderson says. "Her mobilization of young people shows the rising generation expects science to inform policy," she says, "and may inspire many to become scientists themselves."

MICHAEL CAMPANELLA/GETTY

Books & arts



ILLUSTRATIONS BY RICHARD WILKINSON

A book for our time, from all time

Seven leading scientists, historians and scholars choose classic works that speak to now.

From planetary change to geopolitical recalibrations, 2019 has been convulsive. The year saw millions worldwide protesting against governmental inaction on the cascading crises in the global environment. Anxiety over nuclear annihilation vied with concerns over the

repugnant resurgence of 'race science' and the emergent ethics of gene editing. Amid the tumult, *Nature* asked seven scientists, scholars and historians to pluck a book from all time that speaks to our time.

Freeman Dyson, Alondra Nelson, Emilie Savage-Smith, Ann Pettifor, Callum Roberts,

Ismail Serageldin and Chikwe Ihekweazu chose science-inflected volumes – on the often-forgotten lessons of Hiroshima, the rise of unregulated markets, the ubiquity of plastic, mapping the eleventh-century world, and more. Together, they offer a composite lens on our complicated present.

FREEMAN DYSON MESSAGES OF HIROSHIMA

John Hersey's *Hiroshima*, published in 1946 – one year after the destruction of the city by a US atomic bomb – gave the world an enduring vision of nuclear war that has remained the dominant image in the minds of later generations. On the threshold of 2020, the 75th anniversary of the bombing, nuclear war is seen as crowds of half-naked and horribly burnt victims, fleeing from the flames of the burning city, lying down to die of wounds and thirst and radiation sickness. Hersey recorded that scene with unforgettable words in the first part of the book. But that is only half of the US journalist's message.

Hiroshima also shows us an image of nuclear war as a tragedy with heroes as well as victims. Hersey's heroes were doctor Terufumi Sasaki and Methodist pastor Kiyoshi Tanimoto. Sasaki worked with barely a break for three days and nights, using whatever bandages and medicaments he could find in the wreckage of his hospital, easing the pain and hoping to save the lives of an unending stream of sick and dying people pouring in from the surrounding ruins. Tanimoto ran through the burning city to Asano Park, where thousands of victims covered the ground. Soon, the flames advanced across the park. He found a boat on a nearby river and spent a day ferrying sick and dying people to safer ground. He stayed in the park for five days and nights, organizing teams of able-bodied people to bring food and cook meals for the wounded.

The two never knew how many lives they had saved. Each certainly saved several hundred.

After the first days of horror and heroism, Hersey shows the destroyed city coming back to life, with fresh green grass and wild flowers quickly covering the ashes. For weeks, people who have been within a kilometre and a half of the explosion are dying of radiation sickness. After a month, those who are still alive slowly recover. After two months, the survivors are mostly back at work. The city is reborn as a community, with rich and poor sharing the hardships, and widows, widowers and orphans starting new lives.

Hiroshima ends with a quote from an essay written a year later by Toshio Nakamura for his teacher at school. Nakamura was ten years old when he lived through the disaster at Asano Park. "The neighbors were walking around burned and bleeding," he wrote. "We went to the park. A whirlwind came. At night a gas tank burned and I saw the reflection in the river. We

stayed in the park one night. Next day I went to Taiko Bridge and met my girl friends Kikushi and Murakami. They were looking for their mothers. But Kikushi's mother was wounded and Murakami's mother, alas, was dead."

The second half of Hersey's message is that we are a tough species, evolved to survive all kinds of calamities, including the calamity of nuclear war. Individuals die, but communities survive. Unfortunately, the public heard only the first half: the picture of doom. The response was to rush into a frenzy of bomb-building that made the dangers of nuclear war a hundred times worse.

Had we heard the whole message, we would perhaps have chosen a wiser course: saying no to nuclear weapons as we have said no to biological weapons, building a saner world with manageable risks.

Freeman Dyson, retired professor in the School of Natural Sciences, Institute for Advanced Study, Princeton, New Jersey, on **Hiroshima** John Hersey Alfred A. Knopf (1946).

ALONDRA NELSON THE RETURN OF EUGENICS

Some three decades ago, as global concern was tuned to the reunification of Germany, Nelson Mandela's release from a South African prison and the launch of the Human Genome Project, sociologist Troy Duster published a quiet but prescient primer for the dawning age of DNA.

Backdoor to Eugenics foretold a world in which the power of genetics extends well beyond its therapeutic potential. In this world, genetic explanations are offered for issues better explained by politics and social structure, such as inequality; the impact of genetic screening programmes depends on the resources of patients; and the state and businesses fund genetic testing, assembling large caches of personal data, with high stakes for the medical and criminal-justice systems.

These developments did materialize, laying the cornerstone for the social science of genetics. The book also anticipated some of the thorny ethical and political questions we face in today's post-genomic era. For instance, how did screening of newborn babies for medical conditions – once seen as radical, now ubiquitous in the United States – become legitimized? Do we own our own DNA data – and should they be readily available to clinical researchers and the police without oversight?

In the new genetics, Duster argues,

human life is seen through a narrow lens (a phenomenon he calls the "prism of heritability"). He uses the metaphor of front and back doors to illuminate cases in which that lens is used, respectively, in overt or covert ways. And he shows how genetics slides into eugenics.

For Duster, eugenics through the "front door" was exemplified by Nazi Germany's mobilization of science, technology, propaganda and statecraft to demonize people with traits deemed 'unfit'. That culminated in the genocide of Jewish people, people with disabilities and others. There were echoes of this in the United States, with exclusionary immigration policies and the forced sterilization of people including those considered 'feeble-minded'.

Lessons were learnt in the wake of this reprehensible history, and by 1990 it seemed as if scientific racism was highly unlikely to repeat itself. Duster noted that the front door to eugenics was effectively closed. However, the back door was at risk of being opened: emergent thought and practice formed a slim 'wedge'. This insidiously associated genetics with 'target' or 'at-risk' populations that mapped onto vulnerable and marginalized communities.

Once this wedge was inserted, Duster argued, a whole infrastructure of surveillance could be set in place. It might begin with turning some voluntary genetic testing into social requirements. (The mandatory sampling and analysing of DNA from certain criminal suspects in many US jurisdictions is a case in point, as is genomic surveillance in China.) It might involve increasing what was tested, from acute medical conditions to social phenomena such as 'educational attainment'. With new norms in place, data collection and scrutiny could be expanded.

Duster's "backdoor" genetic analyses share qualities with today's direct-to-consumer DNA testing. Both are voluntary, tout the value of individual and community participation, and seem to be benign agents of self-knowledge and health information. But the issue of data misuse and surveillance is now unavoidable.

The increasingly crucial message of *Backdoor to Eugenics* is that genetic disorders and social orders are inextricably linked. Duster made a provocative argument about the way in which intertwined political, cultural and technical forces were giving rise to broad and potentially dangerous uses of genetics. He cautioned, for instance, that the seemingly benign medical surveillance of newborn screening risked diverting attention from challenges that confront black and low-income mothers in the United States, such as access to high-quality prenatal care.

Books & arts

Duster's concerns have been borne out in many ways – for instance, the growing acceptance of genetic scrutiny of embryos through processes such as fetal DNA screening for Down's syndrome. More dramatically, an untested gene-editing procedure was used last year to make heritable changes to the world's first 'CRISPR babies'.

Then, as now, key questions remain. What kind of genomics should we have, and for whom? What assumptions are being made about patient populations, ported into genetics analysis and reified as research findings? In this moment, the back door seems more gaping than ajar.

Alondra Nelson, Harold F. Linder Chair at the School of Social Science, Institute for Advanced Study, Princeton, New Jersey, on **Backdoor to Eugenics** Troy Duster Routledge (1990).

EMILIE SAVAGE-SMITH MAPPING UNCERTAINTIES

Despite vast advances in many fields of science, we live with uncertainties, from the progress of climate change to the nature of consciousness. As we try today to understand and explain the world, a book written 1,000 years ago can still speak to us – not least in its concern with predicting everything from flooding to war. In *The Book of Curiosities of the Sciences and Marvels for the Eye*, a well-educated Egyptian tried to draw together everything he could learn about the structure of the heavens and Earth.

His precise identity is unknown, but he had a love of maps and diagrams. In these, beginning with the fixed stars and Saturn – the outermost planet visible to the naked eye – he worked his way down to Earth's surface, with its vast oceans surrounding landmasses occupied by peoples of varying appearances and customs.

Compiled between 1020 and 1050, *The Book of Curiosities* exemplifies the intense intellectual voraciousness of scholars in Egypt during the era. Cairo was then the centre of a global maritime power, with tentacles stretching from the eastern Mediterranean to the Indus Valley and down the East African coast. The city boasted some of the best-known figures in the history of Islamic science, including the highly original observational astronomer Ibn Yunus, and Ibn al-Haytham (Alhazen), renowned today for his work on optics. The book is preserved in a richly illustrated Arabic manuscript acquired in 2002 to mark the 400th birthday of the Bodleian Library in Oxford, UK. Over the subsequent decade, it was fully translated and analysed.

With neither telescopes nor microscopes to help them, scholars of the early medieval Arabic-speaking world tried to make sense of the Universe and the constantly changing seas



and lands around them. As they knew from the scientific writings of the day, Earth was without question spherical. Various attempts were made to calculate its circumference. The ninth-century caliph al-Mamun sent astronomers into the desert of what would become Iraq to determine the length of a meridian arc of one degree. The eleventh-century Persian scholar al-Biruni applied trigonometry to the problem – as well as tackling the relative size of the five planets visible to the naked eye.

All events in the skies were thought to affect those on Earth; the microcosm mirrors the macrocosm, it was believed. So celestial mapping became something of an industry, not only for timekeeping but for the prediction of earthly events. The ability to forecast winds, earthquakes, storms, droughts, famines and wars was of great importance in that vulnerable era. Some concerns of our medieval Egyptian author are strikingly similar to ours. When discussing the futility of trying to portray on a map any coastline in precise detail, he says: "Sometimes the lower parts of a region are inundated, and we have witnessed in our short lifetime wastelands and passable lands overcome by sea."

Most of his correlations and explanations have been long rejected. In reading *The Book of Curiosities*, however, you cannot but have respect for the observational skills on show, as well as for his logical reasoning. For instance, he argued that the Nile floods were the result of snow melting on equatorial mountains. The work is a much-needed reminder that we are not the only intelligent people to have inhabited this planet. And it makes you wonder: what

will generations 1,000 years hence think of our scientific theories and explanations?

Emilie Savage-Smith, a retired professor of the history of Islamic science at the University of Oxford, UK, on **The Book of Curiosities of the Sciences and Marvels for the Eye** Anon (1020–1050).

ANN PETTIFOR THE MARKET FOR DYSTOPIA

We live in turbulent and uncertain times. Political insurgencies have erupted from Santiago to Hong Kong. Citizens have risen up in anger against ruling elites. Institutions trusted for upholding democracy, the law and public discourse are undermined daily by political leaders on both sides of the Atlantic. More and more public space and wealth are being privatized. Everywhere there is fear: of financial and economic collapse, of a loss of national identity and sovereignty, of political upheaval, of trade wars and real wars. And there is a growing fear that our ecological life-support systems are poised to collapse. These fears, along with the marketization of society, fuel the rise of protectionism, nationalism and even authoritarianism.

Are we witnessing the dissolution of economic globalization, the international system on which Western prosperity and political stability has depended for more than 40 years?

For an understanding of the forces at play, there is no text more illuminating

than Karl Polanyi's 1944 classic, *The Great Transformation*. In this, his most famous book, Polanyi sought to explain the economic, social and political forces that led to the twentieth century's catastrophic world wars and the march of fascism.

Polanyi – an Austro-Hungarian economic historian and social philosopher – noted that nineteenth-century society rested on two pillars: liberal capitalism and representative democracy. Liberal capitalism, in turn, rested on the gold standard, a system of governance that embraced world markets in capital, currencies and commodities. Thus, both domestic and international markets were effectively governed by private, not public, authority. Governments were gradually stripped of autonomy in key economic policy decisions. This internationalized market system demanded that society be subordinated to its needs, argued Polanyi in his lectures (see go.nature.com/2pajnpd). Markets then became detached from political systems of regulatory democracy, which were necessarily bound by borders.

For Polanyi, that separation was the system's deep flaw and “the clue to its rapid downfall” in 1933. The idea of self-adjusting international markets, detached from societal regulation and oversight, implied a bleak utopia indeed. Such an institution could not exist, he argued in *The Great Transformation*, without annihilating the human and natural substance of society. The

gradual changes leading up to its dissolution in 1933, as part of US president Franklin Delano Roosevelt's response to the Great Depression – were in progress long before the start of the First World War, Polanyi explains. But they remained unnoticed at the time. In a reflection apt for our times, he notes that “a society does not become conscious of the true nature of the institutions under which it lived until those institutions have already passed”.

Today, Big Oil, Big Tech and Big Banks effectively police themselves. They have moved “from offering utopia to selling dystopia”, as economic analyst Rana Foroohar argues (see go.nature.com/3822vkb).

The effective organization of the world today is economic, not political. As Polanyi predicted, citizens are belatedly discovering that their politicians and political institutions are impotent against these forces. His book is truly one for our times.

Ann Pettifor, director of Policy Research in Macroeconomics (PRIME) in London, on **The Great Transformation** Karl Polanyi Farrar & Rinehart (1944).

CALLUM ROBERTS DAWN OF THE PLASTIC WORLD

Books that predict the future, particularly one 70 years distant, are usually memorable for how far-fetched or quaint their prophecies seem today, not for their veracity. *Plastics*, a slender volume published in 1941 by two British chemists, Victor Yarsley and Edward Couzens, is exceptional in a different way. It is both uncannily prescient and marred by an enormous blind spot.

Yarsley and Couzens were at the forefront of the plastics revolution, making rapid advances in the field of polymer chemistry and the manufacture of original products from these near-miraculous new materials. Plastics, they wrote, cannot corrode and are sturdy, lightweight and “of a clarity exceeding that of glass if required”. Good insulators, the materials are pleasant to the touch and exceptionally resistant to acids and oils. “The manufacturer of the future will say, not ‘of what material shall I make this article?’ but what kind of plastic shall I use?” they declare.

Taking this as their lodestone, Yarsley and Couzens describe the future for ‘Plastic Man’. For children, it will be a world “of colour and bright shining surfaces”, almost unbreakable, safely rounded and easily cleaned. In this polymer utopia, the growing child “cleans his teeth and brushes his hair with plastic brushes with plastic bristles”, wears “synthetic silk and wool fastened with plastic zip fasteners” and sits on moulded plastic furniture. In old age, plastic dentures and plastic-lensed

spectacles beckon. The material will, in short, be ubiquitous.

Yarsley and Couzens conclude their homage by writing that when the smoke and mess of the Second World War have dissipated and the world begins to rebuild, the return of a newly powerful, industry-driven science will lead to “a new, brighter, cleaner and more beautiful world”. That line jars today. In fulfilling their utilitarian promise, plastics have become a blight of modern life, invading soils, waterways, seas and even the atmosphere. Some 150 million tonnes of plastic circulate in our oceans now, and in the United States alone, more than 26 million tonnes reached landfill in 2017.

One of the great lessons of history is that scientists, in running away with their enthusiasms, perceive the consequences of their inventions only selectively. There is not a word in the book on plastic waste. Perhaps the writers thought plastics would last forever. Nor is there a single sentence on recycling; that is odd, in the Second World War world of ‘make do and mend’, cooking from scraps and saving cardboard. The idea of recycling was gaining traction in the oil industry of the time, but perhaps it didn't fit with the authors' vision of a shiny new consumer paradise. Ultimately, if Yarsley and Couzens had had the vision, their prophecy of a plastic world might have extended to the tide of plastic waste now choking the planet.

The biographical sketch of Couzens provides one of the best lines of the book: “Though he is a firm believer in the future of plastics, he himself prefers glass and metals.”

Callum Roberts, professor of marine conservation at the University of York, UK, on **Plastics** V. E. Yarsley and E. G. Couzens Pelican Books (1941).

ISMAIL SERAGELDIN A MIND FOR RADICAL EQUALITY

We are challenged today by the results of our past actions and current lifestyles. They are complex and tangled: climate change, biodiversity loss, water shortages and regions beset by both swift population growth and potential famine. We are in a battle to live sustainably.

Some might say that this is a moment for English cleric Thomas Malthus's 1798 *An Essay on the Principle of Population*. Yet that book assumes that humans are no different from animals, and respond to resource availability in the same way. I look to a very different work by a contemporary of Malthus's: the remarkably optimistic *Sketch for a Historical Picture of the Progress of the Human Mind*. Its author, the mathematician and philosopher



Books & arts

Nicolas de Caritat, Marquis de Condorcet, saw no limits to the capacity of human intelligence, and called on his readers to use it to build a better society.

Condorcet wrote the *Sketch* while hiding from the extremist wing of the French Revolution. In March 1794, its forces captured him, and possibly murdered him. The essay was published posthumously.

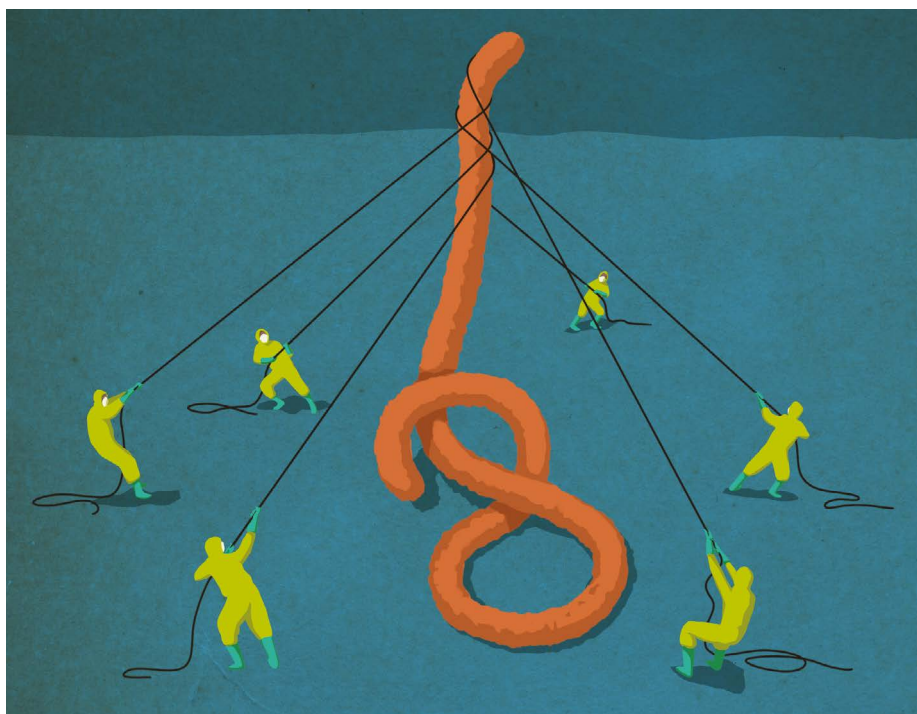
Condorcet had a brilliant and far-ranging mind; his early work included the pioneering *Essay on the Application of Analysis to the Probability of Majority Decisions*, in 1785. His world view was a vision of what could – and should – be, if we aspire to a world of reason and respect for our common humanity. His views are striking even for the late eighteenth century, when sociopolitical radicalism abounded. Condorcet advocated the abolition of slavery and equal rights for women, including women's suffrage. He wanted economic freedom, religious tolerance, legal and educational reform. In the 1790 essay *On the Admission of Women to the Rights of Citizenship*, he argued for human rights generated by virtue of our shared intellectual and ethical capacities:

The rights of men stem exclusively from the fact that they are sentient beings, capable of acquiring moral ideas and of reasoning upon them. Since women have the same qualities, they necessarily also have the same rights. Either no member of the human race has any true rights, or else they all have the same ones; and anyone who votes against the rights of another, whatever his religion, colour or sex, automatically forfeits his own.

Condorcet wanted a classless French republic of citizens protecting their freedom through voting. He designed voting systems, including one based on comparative ranking to satisfy majority rule (a method favoured by Nobel-prizewinning economists Amartya Sen and Eric Maskin). At the global scale, Condorcet called for equality among nations. That included improving people's physical health and longevity, education and moral development. But he recognized that the equality he was describing, both for nations and for individuals, is not absolute: it is equality of freedom and of rights.

This remarkable thinker believed that human ingenuity can overcome all obstacles, and that human goodness can steer us away from tyranny and greed. In this time of global challenge and national turbulence, his wise, inspiring ideas deserve to be remembered.

Ismail Serageldin, founding director of the Bibliotheca Alexandrina, Alexandria, Egypt, on *Sketch for a Historical Picture of the Progress of the Human Mind* Nicolas de Caritat, Marquis de Condorcet (1795).



CHIKWE IHEKWEAZU LOCALIZING PUBLIC HEALTH

Peter Piot's *No Time To Lose* is a passionate account of his leading roles in the discovery of Ebola, the most consequential emerging disease of this decade, and in the global response to HIV and AIDS. I find it speaks profoundly to the current situation in Africa.

As a public-health epidemiologist, I have grown up professionally in the era of AIDS. I have visited Yambuku in the Democratic Republic of the Congo, site of the Ebola virus's first appearance, while supporting the response to a 2004 Ebola outbreak in what is now South Sudan. Thus, *No Time To Lose* felt very immediate to me. Piot draws you in as he describes the appearance of the then-unknown and unnamed Ebola virus in a sample delivered to his laboratory in Belgium in 1976; his first fact-finding trip to Africa; and his professional stint in Yambuku, treating infected people in a hospital run by Catholic nurses. Piot's account of the conversation that led to the naming of Ebola seems almost too simple to be true; despite the virus's severity, there was no naming convention at the time.

In the 1980s, he worked with other scientists investigating many infectious diseases, including HIV infection. As his career pivots to global health politics, he describes in lucid detail his role in the establishment of the Joint United Nations Programme on HIV/AIDS (UNAIDS), and his leadership of the agency between 1994 and 2008. From laboratories to field epidemiology, boardrooms and political chambers, the book charts an incredibly

impactful career in science and the fine arts of diplomacy, communication and political engagement in difficult situations.

The Democratic Republic of the Congo is now grappling with the second-largest recorded outbreak of Ebola, which began in 2018. Despite new tools, more than 3,000 people have been infected and over 2,000 have died in the past year.

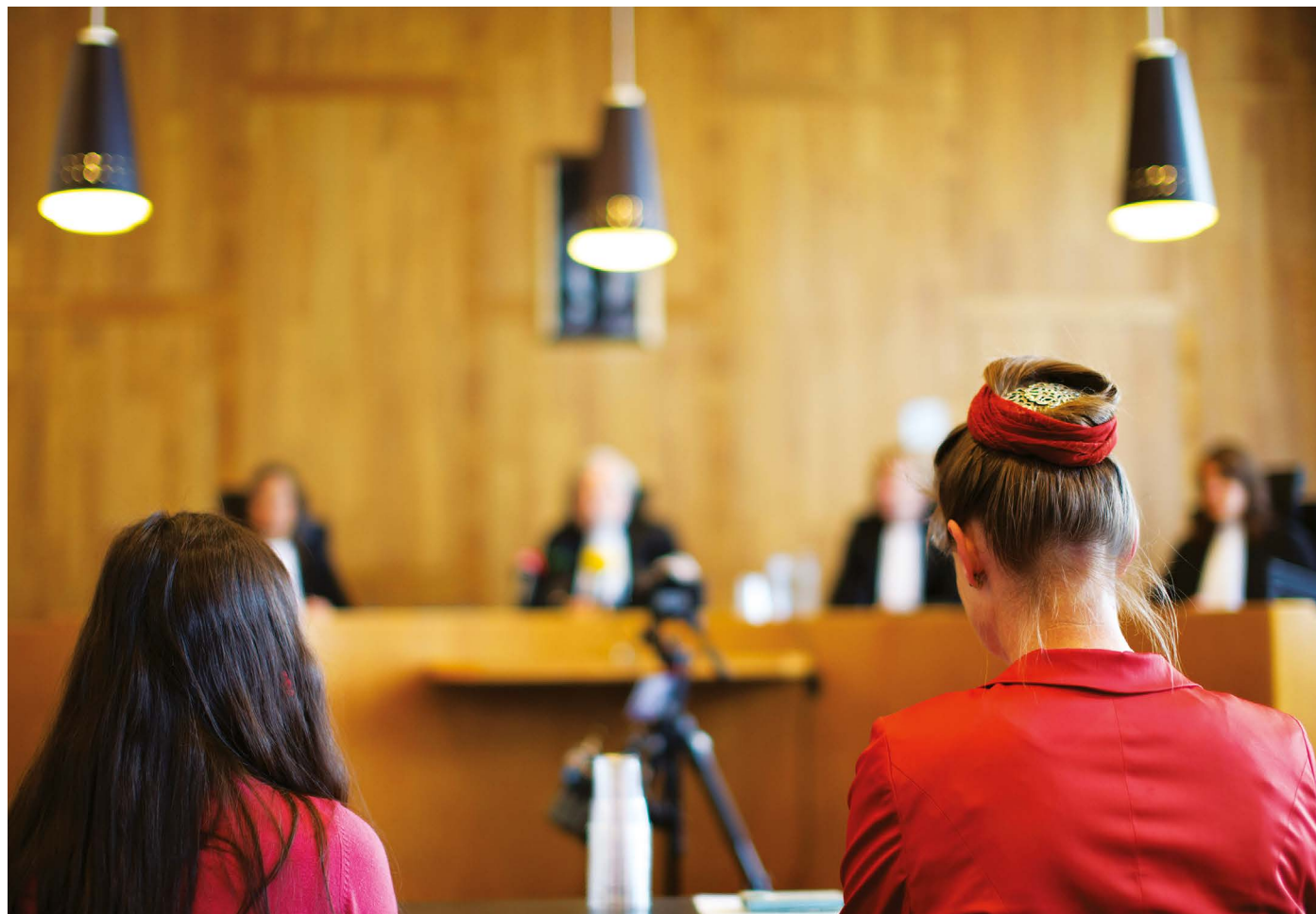
New viruses, such as MERS coronavirus, continue to emerge; old ones persist. At the same time, the population of Africa grows and socio-economic progress has plateaued. As I reflect on the gripping descriptions in *No Time to Lose*, I try to make sense of the paradox of what is happening on the continent.

One of the biggest lessons from Piot's book is that we must focus on building strong, resilient local institutions with a sustainable capacity for infectious-disease prevention, detection and response. Although it was exhilarating to read his account, I caught myself wishing for a different ending: the emergence of a great research institution in Africa. This could train the virologists, epidemiologists and public-health leaders of the future, on the continent where they are most likely to emerge.

A lot of progress has been made since Piot first travelled to Yambuku, more than 40 years ago. As we continue to experience large outbreaks affecting lives and economies in Africa, we must persist in blazing the path towards building local capacity. For that, there really is no time to lose.

Chikwe Ihekweazu, director-general of the Nigeria Centre for Disease Control, Abuja, on *No Time to Lose: A Life In Pursuit Of Deadly Viruses* Peter Piot W. W. Norton (2012).

Comment



PETER DE JONG/AP/SHUTTERSTOCK

The Urgenda Foundation's co-founder Marjan Minnesma (right) and an 11-year-old co-plaintiff wait for a verdict in the Dutch appeals court in 2015.

Not slashing emissions? See you in court

Marjan Minnesma

A pioneer in sustainable innovation explains why she has spent the past decade fighting the first lawsuit to force a government to act on global heating.

I live in a nation where more than one-quarter of the land is already below sea level. For much of the past decade, I've been on a journey for climate justice. With 886 of my fellow Dutch citizens, the Urgenda Foundation that I co-founded brought the first lawsuit aiming to find a national government guilty of failing to safeguard its people from the ravages of climate change. We have won repeatedly, at several levels of the court.

I write this as we await the final ruling in the Supreme Court of the Netherlands in The Hague

on 20 December – a fitting end to a watershed year for civil action on global heating (see the online version of this article for an update on the outcome, at go.nature.com/2pegpu6). The case has inspired other national lawsuits that – along with those against corporations and investors – are creating a burgeoning toolkit of environmental jurisprudence. Together, these serve notice on contributors to the world's still-growing emissions that their inaction is no longer defensible.

In 2011 I read *Revolution Justified* by lawyer Roger Cox (who later acted with lawyer Koos van den Berg for Urgenda in the first court). In the book, Cox argued that catastrophic climate change is a major threat to us and our children, and that governments are not working to prevent it. One of the few democratic ways to make states act, he suggested, is through the legal system.

What if judges read the facts? It would probably be obvious to them that climate change is a clear threat. Might they rule that 'not

Comment

acting' is hazardous negligence that breaches a government's duty of care towards its citizens?

That's certainly how I felt. I had been at the first Conference of the Parties to the United Nations Framework Convention on Climate Change (UNFCCC) in Berlin in 1995. The convention was the focus of my law thesis. In the intervening decades of trying to effect change as a scholar and champion of sustainable innovation, I'd also had three children. With every passing year of empty promises, growing greenhouse gases and rising temperatures, my attitude shifted from cerebral problem-solving to worrying for their future. I now give many speeches, around one-third of which are about the problem and two-thirds about solutions. But, most of all, I like starting projects that seem impossible, and finishing them to leave something concrete.

I decided to bring a lawsuit to force the Dutch government to do what it had said for years was necessary – namely, to reduce the emissions of greenhouse gases by between 25 and 40% by the end of 2020, compared to 1990 levels.

Laggard's letter

You can start a court case only if you have first tried to reach your goals in other ways. So, in November 2012, Urgenda organized a public seminar close to where the parliament of the Netherlands meets, in The Hague. In theory, the parliamentarians who visited could run straight back to the ongoing debate that day and demand of the government what we asked for.

Presenters that day included the outspoken US climate scientist James Hansen, who is now assisting in several court cases brought by groups of young people in the United States and Norway. Another was Urgenda co-founder Jan Rotmans of the Netherlands National Institute for Public Health and the Environment (RIVM). He built the first integrated climate-assessment model, IMAGE, which has been used in international climate negotiations. The audience included politicians, members of the press and engaged citizens, to whom we explained the dangers of doing nothing and the overwhelming evidence of the severe effects of humans' greenhouse-gas emissions on living conditions.

The seminar had little effect.

That month, we wrote a letter to the Dutch government demanding a 40% reduction of greenhouse-gas emissions by 2020. We got a friendly letter back. The government agreed that climate change is a severe problem and that it needed to take action. But, the government wrote, it "did not want to be a frontrunner", claiming that such an approach could dent prosperity and businesses and raise carbon dioxide levels as a result.

This was richly ironic coming from a world-class laggard in sustainable energy. The Netherlands' international reputation for being 'green' is thanks to cycling and recycling. When it comes

to climate change, it talks a lot and does little.

At the time, out of the 27 nations of the European Union, only Luxembourg and Malta generated less energy from renewables than did the Netherlands. Owing to its rich reserves of fossil fuels in offshore natural-gas fields, as well as its massive ports, chemical industries, agriculture and use of coal, the Netherlands was listed 34th of the world's roughly 200 countries in the league table of net emissions that year – more than 80% of all countries emitted less. In the most recent league table, from 2015, it is in 40th place. Looking at the biggest emitters of 2014–16 in absolute terms, the Netherlands was in the top ten for emissions per person, higher than China and way above India.

Suggesting that the nation is 'too small to act', as the state argued in court in April 2015, implies that most countries of the world should also do nothing.

Round one

In mid-December 2012, the Urgenda Foundation decided to sue the government. We invented 'crowd pleading': a cross between crowd funding and citizen science. We asked people to join and help us to look for arguments in court cases all over the world. The foundation gathered the 886 co-plaintiffs, all Dutch citizens, including children – the youngest of whom was 5 years

"People were yelling, crying, applauding and hugging. Hardly anybody had expected we would win."

old when we began. On 20 November 2013, we handed in the summons to the front desk of the Supreme Court in The Hague, demanding a 40% reduction of greenhouse-gas emissions by 2020, or – if this was not possible – at least 25% compared to 1990 levels.

After several rounds of written documents with arguments from us and from the Dutch state, we were called to a hearing at the District Court of The Hague in April 2015. Our hundreds of co-plaintiffs and attendant media could not fit into the court buildings. We produced our own live stream so people could watch together in buildings nearby and follow it at home from their computers. At the end of that day, the judges said they would give their verdict on 24 June 2015 – my 15th wedding anniversary.

We hoped we'd win, but we were not sure at all. We put our chances at perhaps 50%. I never doubted our arguments, but we didn't know whether the judges would have the time and willingness to dive deep enough into the science of climate change.

At 10 a.m. on 24 June we were again in court, to hear the short summary of the three judges on our case. I sat at the front of the room watching the judges and trying to tweet the main

conclusions. Halfway through the summary, I stopped tweeting because I started to realize that the judges were following our line of reasoning. I glanced at the lawyers to check whether I was right. They were concentrating too hard to catch my eye.

The judges agreed that the Dutch government had breached its duty of care by taking insufficient measures to prevent dangerous climate change impairing the living conditions of its people. They based their arguments on tort law (also called civil law) and the doctrine of hazardous negligence. Because the government had signed many documents from the UNFCCC and the European Union declaring that industrial countries should reduce greenhouse gases by between 25 and 40% in 2020, the judges stated that the Netherlands should at an absolute minimum reduce emissions by 25%. Perhaps 40% is necessary, they declared, but the upper bound is at the government's discretion.

A second after the judges left the court room, it erupted with joy. People were yelling, crying, applauding and hugging. Hardly anybody had expected we would win.

The verdict was announced in Dutch and English simultaneously, which helped to spread the word. In half an hour, the news was all over the world. We were overwhelmed by the reactions. Calls flooded in from people from Canada to New Zealand. Some were crying on the phone, saying that they had almost given up, but now had hope again.

For Urgenda, the court case changed a lot. Begun in 2007 at the Erasmus University in Rotterdam, the foundation (now based in Amsterdam) had been a non-governmental organization that mainly worked on solutions to climate change for the Netherlands. In 2008, for instance, we imported the first electric vehicles from Norway and sold them to cities such as Amsterdam, while helping to create a network of charging stations. We kick-started the growth of solar power in the Netherlands by organizing the first collective buying initiative in Europe for solar panels and inverters. Our project 'We Want Sun' purchased 50,000 panels, which at the time brought down the prices for rooftop solar installations in the nation by one-third.

After the win, we were framed by journalists and many others as climate activists. They didn't mean it as a compliment. But I took it as one: an activist is one who acts, just as we'd always done. We are still working on climate solutions, but many know us only from the climate case.

Round two

In September 2015, the government lodged an appeal with the court in The Hague, despite a spontaneous international campaign begging it not to – including messages from celebrities such as actor Mark Ruffalo (who has played the Hulk since 2012) and the model Cameron Russell. So began two years in which our lawyer,



Supporters of a US climate-change lawsuit brought by 21 young people joined a rally in Oregon in June 2019.

Koos van den Berg, produced hundreds of pages with more arguments to convince the appeals court. The second verdict came in October 2018, and again we won! All 29 grounds of appeal from the state were declined.

Better still, this day in court was even more damning for our government (and potentially others) than the first. The district court had ruled that the citizen suit could not base arguments on the European Convention on Human Rights because it was brought by an organization (the Urgenda Foundation) rather than by a human – notwithstanding that its co-plaintiffs numbered hundreds of people. The Court of Appeal disagreed. It declared that the Dutch government is obliged, under articles 2 and 8 of the European Convention on Human Rights, to protect inhabitants by reducing emissions by 25% by 2020. So now we had two duties of care, one from tort law and one based on human rights (a higher-order law).

Round three

Shortly after the second verdict, the government appealed again, this time to the Supreme Court of the Netherlands. This court always takes independent legal advice before ruling, normally from one person. In this case, everything was out of the ordinary, so two advisers were called upon: the deputy

procurator general and the advocate general.

On 13 September this year, they delivered their advice: to uphold the earlier judgments. In 80% of cases, the Supreme Court follows the guidance it is given. But this journey has taught us to brace for surprises.

Meanwhile, six years have elapsed since we filed the case calling for action by 2020. Although the 2015 judgment spurred the state to set a more ambitious climate policy for 2030, little was done to meet the 2020 target. The government simply assumed that the judgment would be overturned on appeal. After the second win, that attitude finally changed. To implement the 2020 target, the government has taken measures to close one of the nation's five coal-fired power plants, and has launched new subsidies for energy-saving activities and renewable energy. But with current national emissions reduced by only 15% from 1990 levels so far, a large gap still remains.

To provide a road map for change, Urgenda published a plan on 24 June – the fourth birthday of the first verdict (see go.nature.com/345d4zr; in Dutch). It included more than 700 organizations, including paper manufacturers, farmers, local sustainable-energy co-operatives and large environmental organizations. It set out 40 measures for reducing greenhouse gases by 25% from 1990 levels by the end of

2020. These included driving at 100 instead of 130 kilometres per hour, raising water levels in nature reserves and energy-saving options for the health and industrial sectors. The foundation later added another ten measures.

So there are now 50 ways for the government of the Netherlands to make up for its failure to protect its citizens from warming of more than 1.5 °C. The 700 partners are poised to help, once the government delivers the money and support that are needed.

It has been a long, hard road, with many ups and downs for the whole team, from tense discussions to nights without sleep. But I'm glad we stayed the course and inspired others around the world to say to their leaders: step up.

The author

Marjan Minnesma is co-founder and director of the Urgenda Foundation in Amsterdam, which this year received an honorary doctorate as an institute from the University Saint-Louis in Brussels. She lectures at many universities and is a board member of the energy co-operative OM | new energy in Amsterdam, the Netherlands.
e-mail: marjan.minnesma@urgenda.nl
Twitter: @marjanminnesma

Correspondence

Scientists reflect on a year of civil unrest. Writing from Syria, Bolivia, Sudan, Iran, Chile, Ecuador, Lebanon, Venezuela, Hong Kong and Catalonia, correspondents tell of altered priorities, day-to-day challenges and hope in the dark times.



LUIS ROBAYO/AFP/GETTY

Venezuelan opposition leader Juan Guaidó (second from left) marching with students in January.

MONA FAWAZ LEBANON: CLASSES IN THE STREETS

Lebanon's ongoing financial meltdown and the political dysfunction behind it have fuelled large protests across the country. Since October, everyday life has been disrupted and classes suspended. The challenges are acute because of the country's location in a region of recurrent wars and refugee crises. One flashpoint has been the loss of public spaces in cities, sacrificed to rampant privatization that is turning Beirut into a playground for the rich.

As a researcher into progressive city planning, I consider that my work should be driven by immediate realities.

Despite the daily difficulties, I have discovered a new creativity in the occupation of abandoned theatres, car parks, city streets and public squares,

which serve as forums for open debates about timely topics. We discuss, for example, the significance of public spaces for political transformation and financial schemes that render land and housing unaffordable for urban majorities. I reframe questions, articulate methods and reconsider what is taken for granted. Inspired also by the soup kitchens, free psychiatric clinics and artistic performances that are reclaiming central Beirut, I assigned my students to devising institutional planning mechanisms to support the restoration of the city's historic core as a shared space. My research will continue to document transgressive practices and seek to inform city planners' conception of common good, urban citizenship and collective property ownership.

Mona Fawaz American University of Beirut, Lebanon.
mf05@aub.edu.lb

BENJAMIN R. SCHARIFKER VENEZUELA: SAFETY, THEN SCIENCE

Scientists need freedom and personal safety to work and pursue the truth – not propaganda, ideologies, post-truth politics and alternative facts. As Venezuela enters its third decade of socio-economic and political upheaval, this year's waves of unrest have further obstructed the serious pursuit of science in my country. This will change only when the complex humanitarian emergency afflicting us today is resolved.

Scientific activity in Venezuela expanded during the second half of the twentieth century, when the country enjoyed relative political and economic stability. Migratory inflows, mostly from Europe and Latin America, favoured the emergence and consolidation

of academic institutions. As the authoritarian pretensions of Hugo Chávez's regime took hold, I became involved in the management of two important universities – the Simón Bolívar University and the Metropolitan University, both in the capital, Caracas – in an attempt to build on this research base. The regime of Nicolás Maduro is now close to achieving its goal of disbanding academic research. The Venezuelan economy has shrunk by two-thirds in the past four years. Shortages of electricity, water, food and medicines have driven around 13% of the population out of the country – the largest refugee crisis in the history of the Americas. My electrochemistry laboratory limps on with just a handful of students.

Benjamin R. Scharifker
Metropolitan University, Caracas, Venezuela.
bscharifker@unimet.edu.ve

MÓNICA MORAES R. BOLIVIA: CREATIVITY FELL TO ZERO

After 14 years of increasingly authoritarian government, a disputed election plunged Bolivia into crisis this autumn. Amid protests, strikes, violence, vandalism, and shortages of fuel and food, research was suspended for more than a month.

With such uncertainty, concentration and abstraction stand no chance. Between silent days and those punctuated by explosions, we put the planning of classes and field trips on hold to keep students and support staff safe. Opportunities were lost for gathering data on our biodiversity, and still need to be rescheduled. Distraction was total, creativity fell to zero and research papers lay unwritten.

My greatest hope is that normality will soon return, particularly to everyday science, so that we can rebuild our confidence and country with a new vision.

Mónica Moraes R. University of San Andrés, La Paz, Bolivia.
mmoraes@fcpn.edu.bo

MUNTASER E. IBRAHIM SUDAN: THE LABS LIE EMPTY

Sudan's political crisis, triggered by government austerity measures imposed a year ago to fend off economic collapse, has sunk the country's mechanisms for learning and research into recession. University gates are shut and laboratories lie empty. The country's young professionals – including doctors, teachers, lawyers, university staff and students – consider this a fair price to pay as they call for a fresh beginning that they demand

should be based on the triad of revolutionary concepts: freedom, peace and justice.

Last summer, this peaceful pro-democracy group brought president Omar al-Bashir's repressive 30-year regime to its knees. The ousted government had inaugurated its reign by launching an anti-science campaign (see *Nature* 348, 5; 1990). One example was

“The ousted government had inaugurated its reign by launching an anti-science campaign.”

the notorious persecution of Farouk Ibrahim, a professor at the University of Khartoum, for teaching evolutionary theory. The latest protests have helped to redeem Ibrahim and end the system of corruption.

Muntaser E. Ibrahim University of Khartoum, Sudan.
mibrahim@iend.org

SHARIF MORADI IRAN: WE CAN RISE ABOVE SANCTIONS

In November, Iran experienced nationwide protests against a government decision to ration petrol and raise its price. Sanctions are among the factors blamed for fuelling the nation's economic and other woes (see, for example, *Nature* 574, 13–14; 2019). Although sanctions since 1979 have been punishing, they have helped to promote homegrown scientific enterprises. For example, Iranian researchers are now producing pharmaceuticals (such as stem-cell treatments for blood diseases, and recombinant chemotherapy drugs), biological research materials and diagnostic kits.

Instead of pinning its hopes on international negotiations, I believe that the government should be focusing on the immense potential of its highly educated people and its plentiful natural resources for making products that are in short supply.

Successive governments in Iran have expressed strong interest in research and development, but they need to spend much more on research to realize its potential – currently this amounts to roughly 0.6% of gross domestic product.

To move forward, Iranian researchers must cultivate their social capital. They should design joint projects with their peers, exchange students with other countries and strengthen connections between academia and industry. When a nation is subject to economic pressure, projects should address crucial local needs. And people should be encouraged to donate money to fund science.

Sharif Moradi Royan Institute for Stem Cell Biology and Technology Tehran, Iran.
sharif.moradi@royaninstitute.org

ERNESTO MEDINA ECUADOR: FOOD DRIVE KEPT CAMPUS OPEN

Academia in Ecuador, already under stress in our struggling economy, was dealt another blow by the civil unrest in October 2019 over more austerity measures. It reportedly caused the country losses of tens of millions of US dollars. Most universities were hit by the protests, which paralysed travel and resulted in shortages of food, fuel and medical supplies for almost two weeks. Barricades allowed no access to labs, so experimental research ground to a halt.

My own institution, Yachay

Tech – one of Ecuador's first research-intensive universities – was protected, however, because most students and many staff members live on campus. The authorities, faculty and administrative employees set up a food drive through temporary barricade openings.

Better still, it has been announced that our budget will be increased for 2020, so there will be new job opportunities for researchers. Orders for research equipment have finally come through.

In spite of all the difficulties, we are now in a good position to consolidate the university's standing in research: Yachay Tech is currently at the top of the Nature Index for Ecuador and among the top 20 institutions in the production of Scopus-indexed articles in the country.

Ernesto Medina Yachay Tech University, Urcuqui, Ecuador.
emedina@yachaytech.edu.ec

WASIM MAZIAK SYRIA: ADVERSITY SOWS RESILIENCE

As the director of the Syrian Center for Tobacco Studies, the civil unrest in Lebanon this autumn reminded me of the upheaval that rocked my home country in 2011, which forced me to move all my research from Aleppo to Beirut. The growing instability in the region amplified my doubts about investing more time in collaborative research, and highlighted the irony of planning research in countries where all aspects of life are hijacked by corrupt and authoritarian regimes. But I pressed on, branching out in my research from a focus on tobacco control to topics related to humanitarian needs.

To do that amid the extreme cruelty of war is not easy.

Correspondence

It demanded that I dig out all the resilience strategies that I have learnt working in unstable circumstances. Relying on local scientists and distant mentorship, building contingency plans and choosing efficient study designs are some of the tips I can offer here. I am also reminded that humanitarian research is critical to bring the suffering of ordinary people, trapped in these painful twists of history, to the attention of the outside world. The hope is that once representative political systems are in place, these seeds will jump-start the collaborations needed to steer effective health solutions.

Wasim Maziak Robert Stempel College of Public Health and Social Work, Florida International University, Miami, USA. wmaziak@fiu.edu

MAI HAR SHAM HONG KONG: KEEP COLLABORATING

The unprecedented social unrest in Hong Kong that was sparked by this year's anti-extradition bill has been going on since June. Last month, it found its way into universities: protesters occupied campuses, staff and students were forced to evacuate, and classes were suspended.

Many students fled. Research laboratories were abruptly shut down, experiments came to a halt, and animal and other facilities were closed. I feared that this situation would drag on, but for my university the disruption was fortunately short-lived.

Hong Kong's scientific research has always been conducted by a complement of local and international researchers. The disturbances have affected our recruitment of talent, especially of postgraduate students. Applications from students on the mainland are expected to plummet. Some newly appointed professors are thinking twice about coming to



A journalist works amid items left by protesters at Hong Kong Polytechnic University in November.

ADNAN ABIDI/REUTERS

join us in Hong Kong.

Universities should be sites of innovation, where we find solutions for problems – including those of socio-political systems. My university identified the SARS coronavirus in 2003, and our research remained strong despite the global financial crisis in 2008 and the Umbrella Movement protests in 2014. I remain optimistic that our extensive scientific research collaborations with academic institutions in China will continue in spite of the current upheavals.

Mai Har Sham University of Hong Kong, Hong Kong, China. mhsham@hku.hk

CECILIA HIDALGO CHILE: SCHOLARSHIP IS KEY TO EQUITY

It is hard to argue for stronger support for science in Chile's current situation of civil unrest, sparked by this year's legitimate protests for social justice that have led to shocking human-rights violations and deplorable violence (see *Nature* 575, 265–266; 2019). Yet critical situations can present an opportunity for improvement. As president of Chile's Academy of Sciences, I contend that more investment in research will

help in the understanding and correction of social injustices and will accelerate the country's long-term development.

To achieve an equitable society, Chile needs to advance the generation of knowledge in all areas – including the natural and social sciences, the arts and the humanities. It must urgently address water scarcity arising from the climate-change-driven desertification of much of the country; the childhood obesity epidemic; the challenges of dealing with a population that is ageing at developed-world rates but with developing-world health care; and the factors fuelling the current social unrest.

These problems require science-based solutions. That means markedly increasing Chile's funding of science, technology and innovation from its present meagre level of 0.36% of gross domestic product.

Cecilia Hidalgo Chilean Academy of Sciences, Santiago, Chile. chidalgo@med.uchile.cl

JOAN MARTÍNEZ ALIER CATALONIA: SCIENCE RISES ABOVE RIOTS

October saw riots in Catalonia – the latest of several waves of demonstrations that have taken place over the past

two years demanding complete autonomy from Spain. Many university students were on strike, but like other researchers, I could still access my offices.

In my view, Catalonia's history of unresolved political tensions has not damaged its recent progress in science. The foundations seem strong: for example, the Catalan Institution for Research and Advanced Studies has contributed substantially to the quality of new research institutes and some university departments since 2001. Moreover, Catalonia holds fourth place – ahead of Spain – among member countries of the European Research Area for the number of European Research Council grants per million inhabitants. (see go.nature.com/2pgfbcd).

I believe that investment in Catalonia's science could increase considerably should independence be obtained from Spain after an agreed referendum. One reason is that fiscal transfers to Spain, which amount to about €2,000 (US\$2,226) per capita each year, would stop. Another is that pro-independence parties explicitly believe in the potential competitive advantage of science in Catalonia.

Joan Martínez Alier Autonomous University of Barcelona, Spain. joanmartineزالier@gmail.com

News & views

Immunology

Identifying the source of tumour-infiltrating T cells

Suman Kumar Vodnala & Nicholas P. Restifo

Immune cells called cytotoxic T cells can recognize and destroy cancer cells. The finding that stem-cell-like T cells exist in tumours, at niche sites that support these cells, could aid efforts to boost anticancer immune responses. **See p.465**

Certain anticancer treatments have been revolutionized by the ability to harness a person's own immune cells for therapeutic purposes¹. Such immunotherapy can result in lasting anticancer responses in people with advanced-stage blood cancers or solid tumours. But not everyone responds. For a variety of cancers, the presence of cytotoxic T cells – immune cells that can kill cancer cells – in a tumour correlates with, but does not predict, an anticancer response and survival. And it is unclear why robust tumour infiltration by T cells occurs in some people, but not in others. On page 465, Jansen *et al.*² reveal a previously unknown source of tumour-infiltrating T cells.

Because tumour cells can proliferate continuously, tumour-targeting T cells must have a similar ability to persist and divide until the last remaining tumour cell is eradicated. In

people undergoing immunotherapy, a greater longevity of antitumour T cells correlates with a better therapeutic outcome³. Therefore, for effective immunotherapy, it is crucial to understand the factors that influence T-cell persistence and infiltration of tumours. Some clues already exist⁴ about these factors, such as the presence of long versions of structures called telomeres, found at the ends of chromosomes, and high expression levels of the protein CD27 in T cells.

In addition to these factors, another clue came from the identification of a subset of stem-cell-like T cells called memory T cells, which can provide long-lasting immune responses^{5,6}, and which express high levels of TCF7 (previously known as TCF-1). This protein is important for maintaining a stem-cell-like state in T cells that also express the protein CD8 (known as CD8 T cells)^{7,8}. Such

stem-cell-like cells can self-renew and give rise to different types of T cell, including a type of CD8 T cell called a cytotoxic CD8 T cell. The presence of stem-cell-like T cells in people who have cancer was reported previously⁵; however, the anatomical location of these cells had not been elucidated. Jansen and colleagues now show that human kidney tumours contain stem-cell-like T cells that reside in the tumour in niches that support them (Fig. 1).

The authors investigated how the level of tumour-infiltrating cytotoxic CD8 T cells varied. They analysed samples of human kidney tumours obtained from people who had undergone tumour-removal surgery, and noted a wide variation in the level of T-cell infiltration between the samples. In people who had tumours in which CD8 T cells accounted for fewer than 2.2% of cells in the sample, the cancer continued to grow, indicating that the surgery and the person's immune response to the residual cancer cells were insufficient to halt disease progression. By contrast, above this threshold of 2.2% infiltration, cancer growth after surgery was four times slower.

Jansen and colleagues then studied the composition and type of T cell in the tumour samples using a technique called flow cytometry, and identified two distinct sets of T cell. One set consisted of cytotoxic CD8 T cells that express high levels of cancer-killing molecules but that also express 'immune-checkpoint' molecules. Expression of checkpoint molecules can drive cytotoxic T cells to enter a dysfunctional state known as exhaustion, which can occur in the tumour microenvironment after prolonged exposure to cancer cells recognized by the T cells. The

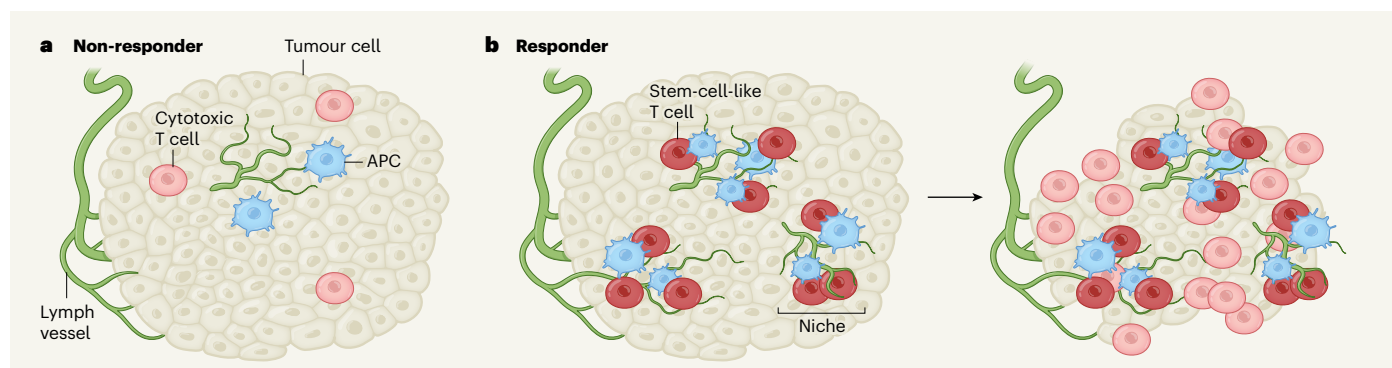


Figure 1 | Stem-cell-like T cells reside in niches in tumours. Jansen *et al.*² report the discovery of stem-cell-like T cells that inhabit kidney tumours. **a**, The authors compared the profile of tumour samples obtained from people who had undergone tumour-removal surgery. In non-responders, whose cancer progressed more rapidly after surgery, there was a low level of tumour infiltration by cancer-killing (cytotoxic) T cells of the immune system. These tumours contain antigen-presenting cells (APCs) and structures

called lymph vessels. **b**, By contrast, people who had longer progression-free survival after surgery had a high level of tumour infiltration by cytotoxic T cells. The authors report that the tumours of these responders had sites, called niches, containing stem-cell-like T cells that could give rise to cytotoxic T cells. These niches were associated with APCs and lymph vessels. There were more lymph vessels in these responding tumours than in the non-responding tumours.

other set consisted of stem-cell-like T cells that Jansen *et al.* demonstrate give rise to cytotoxic CD8 T cells that help to promote an effective antitumour immune response. Stem-cell-like T cells were present only at very low levels in tumours with low levels of T-cell infiltration, whereas tumours with high levels of T-cell infiltration had high levels of the stem-cell-like T cells.

To gain further insight, the authors assessed cellular gene-expression profiles, and analysed epigenetic modifications – types of modification to DNA and its associated proteins that can affect gene expression. They found that, compared with the exhausted cytotoxic CD8 T cells, the stem-cell-like T cells express distinctive immune-signalling molecules called chemokines that are correlated with better patient survival, along with higher levels of key co-stimulatory molecules (which are essential for T-cell differentiation into cytotoxic T cells). Previous analyses^{9,10} of T cells revealed a pattern of progressive steps in epigenetic modification as stem-cell-like T cells give rise to cytotoxic CD8 T cells and then eventually become exhausted.

The epigenetic-modification profile of T cells in tumours can be profoundly influenced by factors in the tumour micro-environment, which can affect the ability of T cells to function as stem cells^{11,12}. For example, the concentration of potassium ions in a tumour modulates epigenetic modifications that influence whether T cells are in the stem-cell-like state that is needed for them to give rise to cytotoxic CD8 T cells^{11,12}. The effect of the tumour microenvironment on the development of cancer-targeting T cells is unclear, and should be a subject for future studies.

Jansen and colleagues noted that the higher than normal expression of chemokines and chemokine-binding receptors in the stem-cell-like T cells is similar to that seen in cells in the microenvironment of lymph vessels – structures through which immune cells move and which support T-cell activation and survival. The authors' analyses demonstrate that stem-cell-like T cells are located in niches in tumours near lymph vessels (Fig. 1), and are confined to dense zones of antigen-presenting cells, which can prime T cells to target tumours. The discovery of these niches by Jansen and colleagues now reveal how stem-cell-like T cells can be maintained in tumours in a functional state capable of generating cytotoxic T cells.

The authors observed a correlation between the presence of protein markers of stem-cell-like T-cell niches and longer, progression-free survival of the people assessed in the study. By contrast, other common ways of assessing an immune response in tumours, such as the expression of the immune-checkpoint protein PD-L1, did not reveal a correlation with progression-free cancer survival.

Previous research¹³ identified stem-cell-like T cells that express rising levels of immune-checkpoint molecules as they progress towards forming cytotoxic CD8 T cells that eventually become exhausted¹⁴. In one example¹³, approaches to block the immune-checkpoint protein PD-1 caused a burst of proliferation in stem-cell-like T cells that express the TCF7 protein. Similarly, in a skin cancer called melanoma, people whose CD8 T cells express TCF7 have a better clinical outcome if they receive immunotherapy to block immune-checkpoint proteins¹⁵. These results suggest that people whose tumours cannot be removed by surgery might benefit from therapy that blocks immune-checkpoint molecules, if their tumours contain stem-cell-like T cells.

Jansen and colleagues' work raises questions about how the stem-cell niches are generated and maintained, and whether tumours might act on them to evade destruction by the immune system. The discovery that resident stem-cell-like T cells exist in specialized niches in tumours suggests that clinical leveraging of such cells to increase the immune infiltration of tumours, together with immunotherapy to

boost exhausted T cells, might unleash T-cell responses to aid the success of anticancer treatment.

Suman Kumar Vodnala and **Nicholas**

P. Restifo are at Lyell Immunopharma, South San Francisco, California 94080, USA. e-mail: nrestifo@lyell.com

1. Rosenberg, S. A. & Restifo, N. P. *Science* **348**, 62–68 (2015).
2. Jansen, C. S. *et al.* *Nature* **576**, 465–470 (2019).
3. Robbins, P. F. *et al.* *J. Immunol.* **173**, 7125–7130 (2004).
4. Rosenberg, S. A. *et al.* *Clin. Cancer Res.* **17**, 4550–4557 (2011).
5. Gattinoni, L. *et al.* *Nature Med.* **17**, 1290–1297 (2011).
6. Gattinoni, L. *et al.* *Nature Med.* **15**, 808–813 (2009).
7. Schilham, M. W. *et al.* *J. Immunol.* **161**, 3984–3991 (1998).
8. Willinger, T. *et al.* *J. Immunol.* **176**, 1439–1446 (2006).
9. Gattinoni, L., Klebanoff, C. A. & Restifo, N. P. *Nature Rev. Cancer* **12**, 671–684 (2012).
10. Crompton, J. G. *et al.* *Cell. Mol. Immunol.* **13**, 502–513 (2016).
11. Vodnala, S. K. *et al.* *Science* **363**, eaau0135 (2019).
12. Eil, R. *et al.* *Nature* **537**, 539–543 (2016).
13. Im, S. J. *et al.* *Nature* **537**, 417–421 (2016).
14. Siddiqui, I. *et al.* *Immunity* **50**, 195–211 (2019).
15. Sade-Feldman, M. *et al.* *Cell* **175**, 998–1013 (2018).

The authors declare competing financial interests: see go.nature.com/2sivhan for details.

This article was published online on 11 December 2019.

In Retrospect

Superconductivity mystery turns 25

N. Peter Armitage

In 1994, an unconventional form of superconductivity was detected in strontium ruthenate. The discovery has shed light on the mechanism of unconventional superconductivity at high temperatures.

Superconductivity is an effect in which a material's electrical resistance vanishes and any magnetic field is expelled below a transition temperature. Despite the remarkable phenomenology, this behaviour is actually quite common: almost half the elements in the periodic table are superconductors¹, albeit at temperatures near or below the extremely low one at which helium gas liquefies (about 4 kelvin). Since Nobel-prizewinning work in the late 1950s, we have had a successful theory² of superconductivity in these conventional systems. Electrons bind into 'Cooper pairs' that have isotropic (direction-independent) properties through an interaction with vibrations of surrounding ions. Over the past 40 years, researchers have looked for unconventional superconductors that involve different pairing

interactions, such as magnetic ones. In 1994, Maeno *et al.*³ reported one of the clearest examples of unconventional superconductivity, in strontium ruthenate near 1 K.

Understanding unconventional superconductors requires identifying both the pairing interaction and the order parameter – a quantity that reflects the interaction and the macroscopic, typically anisotropic, properties of the unconventional superconductivity. The most substantial development in this area of study was the discovery of superconductivity in layered copper-oxide compounds (known as cuprates) in the mid-to-late 1980s. The phenomenon was detected⁴ at the unprecedentedly high temperature (for that time) of 30 K, which led to a worldwide effort to understand the mechanism of cuprate superconductivity.

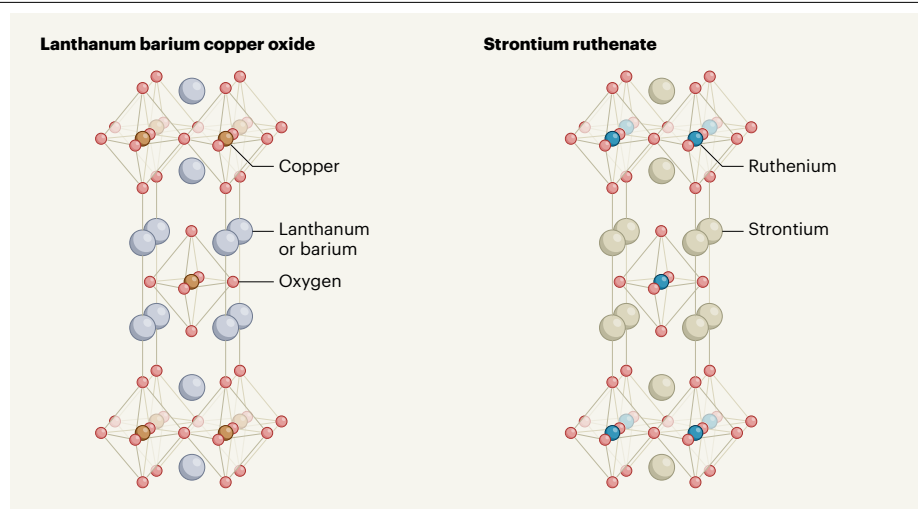


Figure 1 | Crystal structures of two superconductors. In 1986, lanthanum barium copper oxide was found⁴ to superconduct (transport electricity without resistance) at the relatively high temperature of 30 kelvin. Eight years later, Maeno *et al.*³ reported the discovery of superconductivity in strontium ruthenate at about 1 K. Although these two materials have the same crystal structures at high temperatures, their superconductivity mechanisms are likely to be markedly different.

The cuprates are now thought to have a highly anisotropic order parameter, and to have Cooper pairs made of electrons that have anti-aligned spins (intrinsic angular momenta). Such spins form non-magnetic states that have even parity, which means that the wavefunction of the state does not change sign if the signs of the spatial coordinates are flipped. Cuprate superconductivity has been proposed⁵ to arise from an interaction of electrons with antiferromagnetic spin fluctuations (antiferromagnetism is a form of magnetism in which spins are anti-aligned with their neighbours). However, no theory has yet gained general acceptance.

One method that has been used to try to understand these compounds is to search for superconductivity in materials that are related in some way to the cuprates. In this way, it might be possible to identify the structural, electronic or magnetic features that are essential for the materials' high transition temperatures. In particular, the cuprate discovery led to a huge effort to investigate compounds that contain transition metals other than copper.

It was against this backdrop that Maeno and colleagues found superconductivity in strontium ruthenate, at about 1 K. This was decidedly not high-temperature superconductivity. But the work caused tremendous excitement because it described the detection of superconductivity in another layered transition-metal oxide – and in a material that has the same crystal structure as the original superconducting cuprate, lanthanum barium copper oxide⁴ (Fig. 1). Almost immediately, it was realized that there were both similarities and differences between the cuprates and strontium ruthenate.

One main difference is that pure compounds of the cuprates (such as lanthanum copper oxide) are antiferromagnetic insulators and require the substitution of atoms (such as

barium for lanthanum) to conduct electricity. By contrast, pure strontium ruthenate is strongly metallic. A striking aspect of the superconducting cuprates is that their metallic state at temperatures above the transition temperature seems to be even more unconventional than their superconducting state. The metallic state is thought to be the result of strong interactions between electrons. A radically new theory of 'strange metals' might be needed to understand the high-temperature metallic state and thereby also the superconducting state that forms from it⁶. In strontium ruthenate, electron interactions are also strong, but they do not change the fundamental character of the metallic state.

This aspect, and the fact that related materials in the larger ruthenate family exhibit ferromagnetism (a form of magnetism in which spins are aligned with their neighbours), led to the proposal⁷ in 1995 that superconductivity in strontium ruthenate could be an analogue of the superfluid A phase in helium-3. In this phase, the compound exists as a superfluid (a zero-viscosity liquid) made from odd-parity Cooper pairs of neutral helium-3 atoms that have aligned spins⁸. The proposal gained much support, both for the compelling science that suggests it and for the beautiful idea that there could be an odd-parity superconductor driven by ferromagnetism in the same way that the cuprates might be even-parity superconductors driven by antiferromagnetism. Of course, the "great tragedy of Science [is] the slaying of a beautiful hypothesis" by experimental facts⁹. Experiments always have the final say.

The exciting science, the ability to grow large, extremely pure crystals and an exceedingly collaborative research community pushed superconducting strontium ruthenate forward as a highly active topic of

investigation. Moreover, there was the abiding sense that it should be possible to unambiguously determine the nature of the material's unconventional order parameter, because its high-temperature metallic state – unlike that of the cuprates – seemed to obey the conventional theory of metals. This determination is an ongoing saga, with field-changing results coming even this year. Notable early work showed evidence for unconventional odd-parity pairing of electrons in nuclear magnetic resonance (NMR) spectroscopy¹⁰, and for spontaneous generation of magnetism^{11,12} consistent with the proposal outlined above.

In the past five years, sophisticated measurements of strontium ruthenate have failed to show an odd-parity superconducting transition splitting into two under mechanical strain, as had been predicted¹³. These measurements, along with a reinvestigation using NMR spectroscopy¹⁴, have given compelling evidence that the superconductivity is likely to be even parity. But this even-parity state is inconsistent with the experiments that showed the presence of spontaneous magnetism. Therefore, the nature of unconventional superconductivity in strontium ruthenate must be considered unresolved.

This problem, together with that of the cuprates, has pushed theory, experiment and materials synthesis forward in directions that would have been unimaginable when superconductivity in these compounds was discovered. And as is so often the case, many of the ideas that scientists have grappled with in the context of a hard problem have turned out to be incredibly influential in areas well beyond their original scope. In this particular case, important cross-fertilizing connections can be made with topological insulators (bulk electrical insulators that have conducting surfaces) and quantum computation¹⁵. The research community is still hard at work on the mystery of strontium ruthenate. Experiments always have the final say.

N. Peter Armitage is in the Department of Physics and Astronomy, The Johns Hopkins University, Baltimore, Maryland 21218, USA. e-mail: npa@jhu.edu

1. Shimizu, K. in *100 Years of Superconductivity* (eds Rogalla, H. & Kes, P. H.) Ch. 4 (Taylor & Francis, 2011).
2. Bardeen, J., Cooper, L. N. & Schrieffer, J. R. *Phys. Rev.* **106**, 162–164 (1957).
3. Maeno, Y. *et al.* *Nature* **372**, 532–534 (1994).
4. Bednorz, J. G. & Müller, K. A. Z. *Phys. B*, **64**, 189–193 (1986).
5. Nagaosa, N. *Science* **275**, 1078–1079 (1997).
6. Hussey, N. E. *J. Phys. Condens. Matter* **20**, 123201 (2008).
7. Rice, T. M. & Sigrist, M. *J. Phys. Condens. Matter* **7**, L643–L648 (1995).
8. Lee, D. M. *Rev. Mod. Phys.* **69**, 645–665 (1997).
9. Huxley, T. H. *Biogenesis and abiogenesis; collected essays*, vol. 8 (1884).
10. Ishida, K. *et al.* *Phys. Rev. B* **63**, 060507 (2001).
11. Luke, G. M. *et al.* *Nature* **394**, 558–561 (1998).
12. Xia, J., Maeno, Y., Beyersdorf, P. T., Fejer, M. M. & Kapitulnik, A. *Phys. Rev. Lett.* **97**, 167002 (2006).
13. Hicks, C. W. *et al.* *Science* **344**, 283–285 (2014).
14. Pustogow, A. *et al.* *Nature* **574**, 72–75 (2019).
15. Sato, M. & Ando, Y. *Rep. Prog. Phys.* **80**, 076501 (2017).

This article was published online on 9 December 2019.

New antibiotics target bacterial envelope

Marcelo C. Sousa

A double membrane protects certain bacteria from antibiotics, but compounds have now been generated that can overcome this obstacle, seemingly by targeting a crucial protein in the outer membrane. **See p.452 & p.459**

Antibiotic resistance is a growing global public-health problem¹. One group of bacteria, called Gram-negative bacteria, is particularly difficult to treat, because the cells are shielded by a double-membrane envelope, which constitutes a formidable barrier to antibiotics². When antibiotics do breach the membranes, these bacteria often use efflux pumps to remove the drugs^{3,4}. Three papers (two in *Nature*^{5,6} and one in the *Proceedings of the National Academy of Sciences*⁷) now describe antibiotics that overcome these obstacles by targeting, directly or indirectly, a protein integral to the outer membrane.

The outer membrane of Gram-negative bacteria contains lipopolysaccharide (LPS) molecules in its outer leaflet, with outer-membrane proteins (OMPs)⁸ spanning the entire outer membrane. OMPs are folded into the membrane by a protein complex called the β -barrel assembly machine (BAM), the central component of which, BamA, is an OMP itself (Fig. 1). Because BamA is exposed to the extracellular space, it could be an Achilles heel in the bacterial shield – inhibitors that access BamA would not need to penetrate the cell. Indeed, a proof-of-concept study⁹ has shown that this approach inhibits OMP folding and compromises membrane integrity, albeit by an unknown mechanism.

The three current studies took different approaches to develop antibiotics against Gram-negative bacteria. On page 459, Imai *et al.*⁵ turned to Gram-negative bacteria that live symbiotically in the gut of nematode worms and can secrete antibiotics to fend off competing bacteria – including other Gram-negative species. A screen of the secretions from 22 of these symbionts revealed a Gram-negative-targeting antibiotic, which the authors named darobactin.

Darobactin displayed antibiotic activity against multiple Gram-negative bacteria, both *in vitro* and in infected mice, including against several drug-resistant human pathogens such as polymyxin-resistant *Pseudomonas aeruginosa* and β -lactam-resistant *Klebsiella pneumoniae* and *Escherichia coli*. Darobactin as

not toxic to human cells at the concentrations at which it was an effective antibiotic.

Next, Imai *et al.* asked what bacterial molecule darobactin targets. The group identified three strains of *E. coli* that were resistant to darobactin and showed that each harboured mutations in the *bamA* gene. The mutations all changed amino-acid residues in the same region of BamA's protein structure, suggesting a putative binding site for darobactin that would be accessible from the extracellular space.

The authors provided evidence that darobactin and BamA bind to each other directly, using a technique called isothermal titration calorimetry, which measures the heat changes associated with physical interactions between molecules. The results of nuclear magnetic resonance (NMR) spectroscopy experiments

were also consistent with direct binding, and suggested that the antibiotic stabilizes the protein in a potentially inactive conformation.

The researchers next showed that darobactin inhibits the ability of an isolated BAM complex to perform its OMP-folding function *in vitro*, consistent with direct BamA targeting. However, only one of the resistant BamA mutants showed reduced inhibition by darobactin in this assay. A test of whether darobactin–BamA binding is impaired in the *bamA* mutants could be used in the future to confirm BamA as the molecular target.

On page 452, Luther *et al.*⁶ focused on analogues of an existing antibiotic, murepavadin¹⁰, which targets a surface-exposed protein called LptD that is involved in assembling LPSs in the outer membrane⁸. Murepavadin displays potent but narrow antibiotic activity against *P. aeruginosa*¹⁰. The authors therefore screened for murepavadin analogues that had antibiotic activity against other Gram-negative species.

Luther and colleagues chemically linked the compounds identified through this screen to a portion of another antibiotic, polymyxin B, that binds to LPS directly¹¹. Intact polymyxins efficiently disrupt bacterial membranes and kill cells, but are rather toxic to humans¹². The researchers hoped that linking just the LPS-binding portion of polymyxin B could increase the membrane targeting of their murepavadin analogues. Indeed, their strategy produced several chimaeras that had potent activity, both *in vitro* and in mice infected with *K. pneumoniae*, *P. aeruginosa*, *E. coli* and other Gram-negative bacteria,

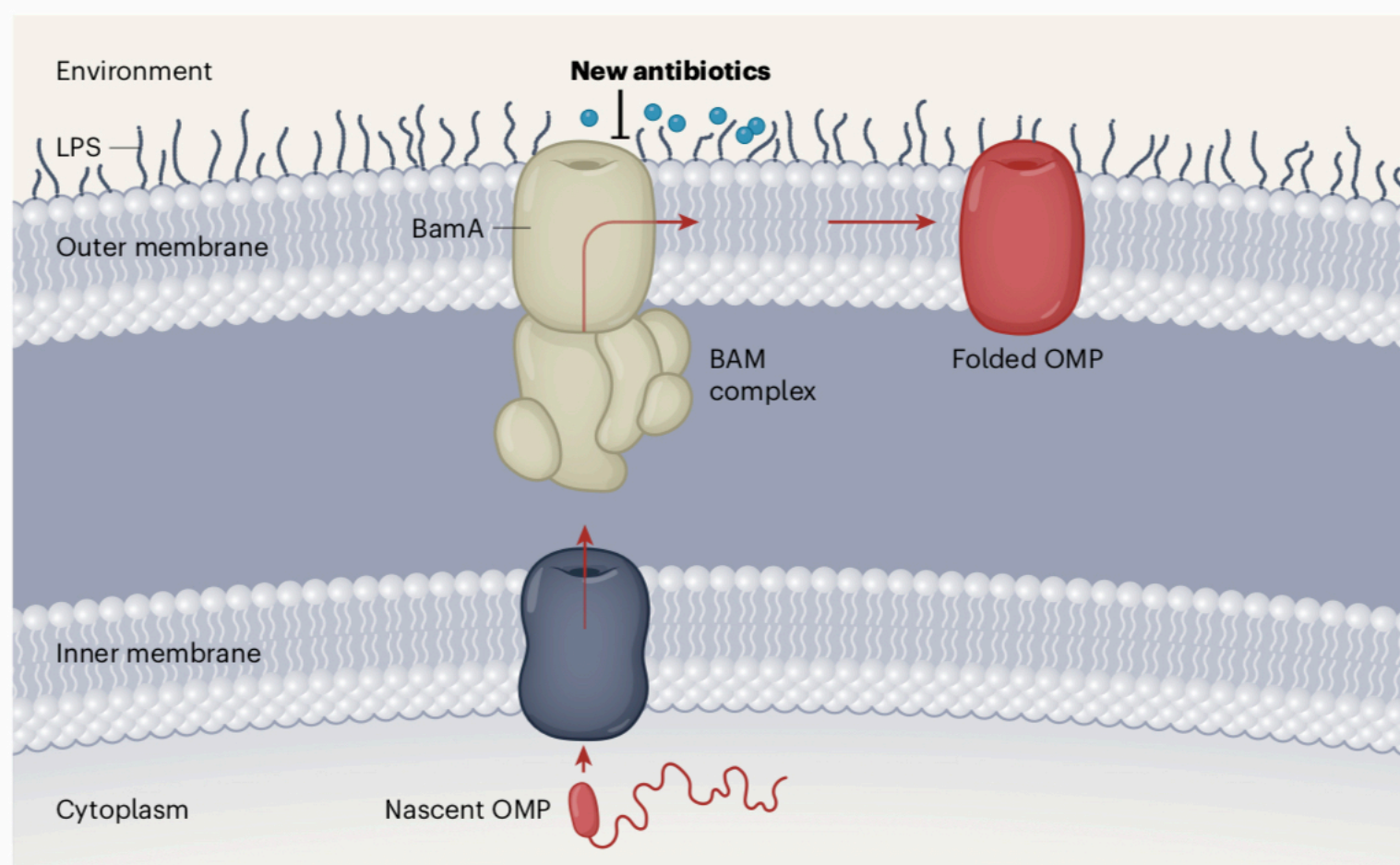


Figure 1 | Overcoming a double-membrane barrier. Gram-negative bacteria are protected by inner and outer membranes. The outer membrane contains lipopolysaccharide (LPS) molecules in the outer layer and integral outer-membrane proteins (OMPs). These proteins are synthesized in the cell's cytoplasm and transported to the space between the membranes by the translocation machinery (dark blue). From here, they are captured, inserted and folded into the outer membrane by the BAM protein complex (red arrows). BamA is the central component of BAM and is accessible from the bacterial surface. Three studies^{5–7} describe new antibiotics that seem to target BamA, preventing the normal OMP folding that is required for bacterial survival.

including drug-resistant strains. Notably, the chimaeras showed low toxicity in mice.

It might be expected that the chimaeras would target LptD, but when Luther and colleagues tested for interacting partners, they found evidence of BamA targeting. The authors analysed strains of *K. pneumoniae* that showed resistance to the chimaeras. They found that resistant strains carried mutations in several genes, including *bamA* and genes responsible for LPS modification. Reintroduction of the wild-type *bamA* gene into the resistant strains led to increased sensitivity to the chimaera, indicating that BamA has a role in the antibiotic's mechanism of action.

Direct chimaera–BamA binding was confirmed with *in vitro* assays in which the authors fluorescently labelled the chimaeras and monitored changes in fluorescence that indicate binding to a large protein such as BamA. As with darobactin, NMR experiments suggested that chimaera binding stabilizes BamA in a potentially inactive conformation, consistent with direct BamA targeting. However, when the bacteria were treated directly with the chimaeras, both the outer and inner membranes were rapidly permeabilized; this suggests that the compounds might act directly on the membrane. The results raise the possibility that the chimaeras act in a similar way to polymyxins, with binding to BamA strengthening their membrane targeting.

In the third study, Hart *et al.*⁷ identified a compound, MRL-494, that had similar antibiotic potency against both wild-type *E. coli* and a mutant defective in outer-membrane integrity and efflux mechanisms, suggesting that this antibiotic might not need to penetrate the cell to exert its activity. *In vitro*, MRL-494 exhibited moderate potency against Gram-negative pathogens, including *K. pneumoniae* and *P. aeruginosa*. The efficacy of MRL-494 in animal models remains to be tested.

The authors showed that treatment of *E. coli* with the compound resulted in decreased abundance of OMPs in the outer membrane, indicating BamA as a possible target. In support of this possibility, Hart *et al.* identified a *bamA* mutation that confers resistance to MRL-494 in *E. coli*. They showed that, whereas MRL-494 inhibited normal folding of a model OMP in *E. coli* cells expressing wild-type *bamA*, it had less effect on the resistant cells. The researchers found that MRL-494 stabilizes BamA against heat-induced protein aggregation in cells, suggesting an interaction between the two. However, MRL-494 stabilizes the resistant *bamA* mutant to a similar extent. Furthermore, MRL-494 displays similar potency against Gram-positive bacteria, which lack BamA. Therefore, in Gram-negative bacteria, MRL-494 might inhibit BamA directly or might target the outer membrane and affect BamA function indirectly.

Together, these studies describe new antibiotics that are active against difficult-to-treat Gram-negative bacteria. Given the compounds' size and chemistry, they are likely to act at the cell surface, bypassing the need to breach the permeability barrier. Imai *et al.* provided compelling evidence that BamA is the target of darobactin, including a putative binding site, to be confirmed by demonstrating

“These studies describe new antibiotics that are active against difficult-to-treat Gram-negative bacteria.”

reduced binding to resistant mutants. The chimaeric compounds both seem to bind BamA and LPS. But, as is also the case for MRL-494, further experiments will be required to determine whether their activity is caused by direct effects on BamA.

Future research to identify specific BamA binding sites for any of the compounds, and to examine the mechanism by which antibiotic binding impairs BamA activity, would provide a platform for further antibiotic development. Such research might also shed light on how BAM mediates the insertion and folding of OMPs, which is poorly understood.

Darobactin and MRL-494 are initial lead

compounds, and medicinal-chemistry efforts could yield more-potent and effective analogues. Preclinical studies aimed at determining their toxicity in animal models will also be important. Luther and colleagues' chimaeras are at a more advanced stage of development, because, as the authors show, they have potent *in vivo* activity as well as favourable toxicity, pharmacokinetics and pharmacodynamics in animal models. The future looks promising for this newly discovered class of antibiotic.

Marcelo C. Sousa is in the Department of Biochemistry, University of Colorado, Boulder, Colorado 80301, USA.
e-mail: marcelo.sousa@colorado.edu

1. World Health Organization. *Global Priority List of Antibiotic-Resistant Bacteria to Guide Research, Discovery, and Development of New Antibiotics* (WHO, 2017).
2. Nikaido, H. *Microbiol. Mol. Biol. Rev.* **67**, 593–656 (2003).
3. Li, X.-Z. & Nikaido, H. *Drugs* **64**, 159–204 (2004).
4. Munita, J. M. & Arias, C. A. *Microbiol. Spectr.* **4**, VMBF-0016-2015 (2016).
5. Imai, Y. *et al. Nature* **576**, 459–464 (2019).
6. Luther, A. *et al. Nature* **576**, 452–458 (2019).
7. Hart, E. M. *et al. Proc. Natl Acad. Sci. USA* **116**, 21748–21757 (2019).
8. Konovalova, A., Kahne, D. E. & Silhavy, T. J. *Annu. Rev. Microbiol.* **71**, 539–556 (2017).
9. Storek, K. M. *et al. Proc. Natl Acad. Sci. USA* **115**, 3692–3697 (2018).
10. Srinivas, N. *et al. Science* **327**, 1010–1013 (2010).
11. Mares, J., Kumaran, S., Gobbo, M. & Zerbe, O. *J. Biol. Chem.* **284**, 11498–11506 (2009).
12. Li, J. *et al. Lancet Infect. Dis.* **6**, 589–601 (2006).

This article was published online on 9 December 2019.

Condensed-matter physics

Magnetic and topological order united in a crystal

Roger S. K. Mong & Joel E. Moore

A material that has electrically conducting surfaces has been found to show, when cooled, a type of magnetic ordering that reduces conduction at the surfaces. Such remarkable behaviour could have practical applications. **See p.416 & p.423**

An ordered arrangement of magnetic moments in a material normally prevents the formation of another kind of electronic order associated with an exotic state of matter known as a topological insulator. But Otkov *et al.*¹ (page 416) and Rienks *et al.*² (page 423) report that manganese bismuth telluride integrates these two types of electronic behaviour. The complex layered structure of alternating magnetic and topologically non-trivial regions in this material leads to an intriguing and potentially technologically useful interplay between magnetic and topological order.

One of the earliest descriptions of electronic order in solids was of ferromagnetism,

the existence of which was reported in natural minerals in Greece and China more than 2,000 years ago. In a simple ferromagnet, microscopic magnetic moments, arising predominantly from the spin (intrinsic angular momentum) of a material's electrons, align in the same direction, leading to an overall macroscopic magnetic moment. The concept of antiferromagnetism, in which spins align in alternating directions and the average magnetic moment is zero, was developed only in the 1930s. These two kinds of magnetic order are viewed theoretically as breaking time-reversal symmetry: when the direction of time is reversed, the pattern of spins

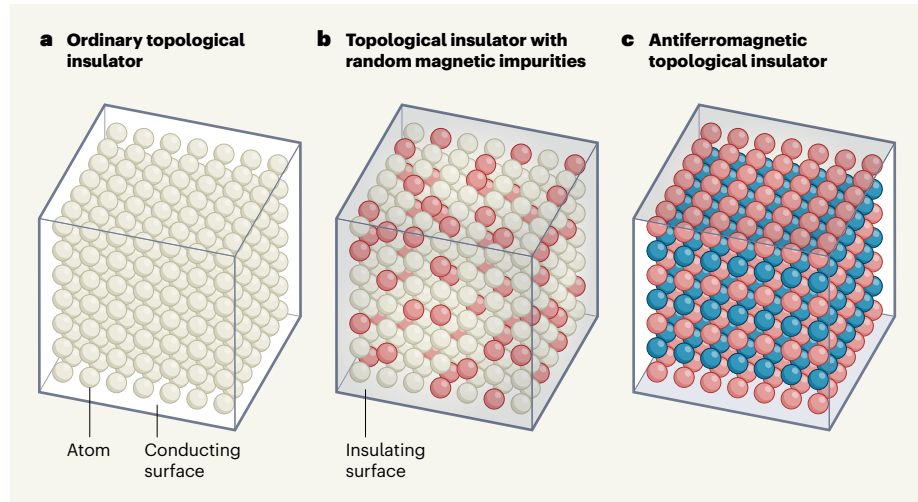


Figure 1 | A new spin on topological insulators. **a**, In an ordinary topological insulator, the atoms are unmagnetized (beige), the interior is electrically insulating and all of the surfaces are conducting. **b**, When random magnetic impurities (red) are introduced, all of the surfaces become insulating. Rienks *et al.*² present evidence that magnetic manganese atoms sit in random locations when added to bismuth selenide, but separate into layers when added to bismuth telluride (not shown). **c**, If the electron spins (intrinsic angular momenta) of atoms form a pattern that alternates between layers (illustrated by the red and blue atoms), an antiferromagnetic topological insulator forms. This has both conducting and insulating surfaces; whether a particular surface is conducting or insulating depends on its orientation relative to the magnetic structure. Here, the top and bottom surfaces are insulating, and the others are conducting. Otrokov *et al.*¹ present evidence for such an antiferromagnetic topological insulator.

is changed. Time reversal acts by reversing velocities (a bit like running a movie backwards) as well as by reversing the direction of angular momenta, including spins.

However, unbroken time-reversal symmetry is required to produce topological insulators, which have been a tremendously active area of study^{3,4}. Topological insulators have an electrically insulating interior, but a conducting surface. The property that sets them apart from ordinary insulators is that their surfaces cannot be made to have a simple insulating state as long as time-reversal symmetry is unbroken. This makes topological insulators ideal platforms for generating excitations known as Majorana zero modes, which are the basis for a topological approach to quantum computing⁵.

The two current papers demonstrate that crystalline manganese bismuth telluride manages to combine these seemingly incompatible magnetic and topological orders. When cooled, the material becomes magnetic, and yet displays a kind of topological insulating behaviour. Unlike an ordinary topological insulator, for which every surface is conducting, such an antiferromagnetic topological insulator has surfaces that are conducting or insulating, depending on how the specific surface cuts through the alternating pattern of spins⁶ (Fig. 1).

Manganese bismuth telluride can be viewed as a stack of magnetic manganese telluride layers that are separated by layers of a benchmark topological insulator, bismuth telluride⁷. Indeed, a thin-film version of this material was produced earlier this year by alternately growing single layers of manganese telluride and

bismuth telluride⁸. The current work shows that crystalline manganese bismuth telluride integrates two of the most actively studied kinds of order: topological insulating behaviour and single-layer magnetism, as seen in monolayer ferromagnets, such as chromium triiodide⁹.

The key feature of the antiferromagnetism in manganese bismuth telluride is that it can retain a modified version of time-reversal symmetry. In an antiferromagnet, the flipping of spins associated with time reversal changes the pattern of alternating spins. However, the combined transformation of first flipping the spins and then shifting the position of the pattern by a unit cell (the smallest repeating unit of a crystal lattice) can leave the pattern unchanged.

Consider adding such antiferromagnetism to a topological insulator. The conducting surfaces of the topological insulator are susceptible to externally applied or intrinsic magnetic fields, because such fields break time-reversal symmetry. Depending on how an antiferromagnetic pattern ends at a surface, the surface can have alternating spins akin to those in a 2D antiferromagnet, uniform spins similar to those in a 2D ferromagnet or a more complicated spin configuration. Consequently, what happens to the surface conduction of the material depends on how these surfaces cut the magnetic order (Fig. 1).

Otrokov *et al.* carried out numerical simulations in conjunction with various experimental probes of manganese bismuth telluride. On the basis of the results, they argue that the material has both an antiferromagnetic order and

a band inversion (whereby the usual ordering of electron energy bands is flipped), which is a signal of a topological insulator. The authors used a method known as X-ray magnetic circular dichroism to confirm the magnetic ordering experimentally.

In addition, these authors used a technique called angle-resolved photoemission spectroscopy to study the electronic structure at the material's surface. They observed that a structure in the surface electron bands, known as a Dirac cone, is modified near and below the temperature at which the material becomes antiferromagnetic. This Dirac cone is the distinguishing feature of the surface of a topological insulator, and its disappearance is a sign that the surface has been converted from conducting to insulating.

Rienks *et al.* carried out detailed studies of the atomic and electronic structure of bismuth telluride to which varying amounts of manganese had been added through a process called doping. They found that a small amount of doping (about 6% manganese) turns the topological insulator into a ferromagnet (as opposed to the antiferromagnet seen by Otrokov *et al.*). Rienks and colleagues also observed a tendency of doped bismuth telluride to form septuple layers, similar to those in crystalline manganese bismuth telluride, but separated by standard, quintuple layers of bismuth telluride. In addition, they studied manganese-doped bismuth selenide, which might have been expected to behave in the same way as its telluride counterpart, given that these compounds are similar topological insulators.

Rienks *et al.* found that manganese added to bismuth selenide does not have as strong a tendency to form layers as it does in bismuth telluride, and they showed that this structural difference has consequences for the magnetic structure and surface electrons. The telluride compound has a gap in energy between surface electron bands and has a magnetization that is perpendicular to the plane of the surface. By contrast, the selenide compound does not have this gap and has in-plane magnetization. The energy gap in the telluride compound disappears when the temperature is raised above the magnetic transition temperature. This connection between the energy gap and magnetism has been challenging to observe in other materials and samples.

A notable requirement for an antiferromagnetic topological insulator is that whether a surface has gaps (insulating) or does not (conducting) depends on the surface termination of the 3D crystal. This difference could be seen by producing other surfaces, or by looking at steps on a surface using scanning tunnelling microscopy. The combination of magnetic and topological behaviour observed in the current papers could lead to new spin-based electronic devices, because the topological aspects

might, for example, improve how materials transport spin currents. There is an overlap in potential applications with a topological phenomenon called the quantum anomalous Hall effect (QAHE), which is generated using magnetic impurities in thin films of topological insulators^{10,11}. Steps on the surface of an antiferromagnetic topological insulator produce the same perfectly conducting edge channels as in the QAHE.

The fact that manganese bismuth telluride is intrinsically magnetic, rather than having its magnetism result from randomly located impurities, as in current QAHE materials, could be advantageous. Key questions remain about how the magnetism varies between different samples of this material. In particular, it seems that both the magnetic transition temperature and the nature of the magnetic ordering between planes could vary, and that an applied magnetic field might be used to switch this ordering. At a more fundamental level, antiferromagnetic topological insulators are predicted to support, without applied electric or magnetic fields, a quantized electromagnetic response (one that comes in discrete units), known as axion electrodynamics⁶. The current papers show

how the synthesis and theory of crystals that have various symmetries combine to reveal important types of electronic order.

Roger S. K. Mong is in the Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15620, USA. **Joel E. Moore** is in the Department of Physics, University of California, Berkeley, Berkeley, California 94720, USA, and at the Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California.
e-mails: rmong@pitt.edu; jemoore@berkeley.edu

1. Otrokov, M. M. *et al.* *Nature* **576**, 416–422 (2019).
2. Rienks, E. D. L. *et al.* *Nature* **576**, 423–428 (2019).
3. Hasan, M. Z. & Kane, C. L. *Rev. Mod. Phys.* **82**, 3045–3067 (2010).
4. Moore, J. E. *Nature* **464**, 194–198 (2010).
5. Fu, L. & Kane, C. L. *Phys. Rev. Lett.* **100**, 096407 (2008).
6. Mong, R. S. K., Essin, A. M. & Moore, J. E. *Phys. Rev. B* **81**, 245209 (2010).
7. Chen, Y. L. *et al.* *Science* **325**, 178–181 (2009).
8. Gong, Y. *et al.* *Chinese Phys. Lett.* **36**, 076801 (2019).
9. Huang, B. *et al.* *Nature* **546**, 270–273 (2017).
10. Chang, C.-Z. *et al.* *Science* **340**, 167–170 (2013).
11. Tokura, Y., Yasuda, K. & Tsukazaki, A. *Nature Rev. Phys.* **1**, 126–143 (2019).

remarkably successful at treating leukaemia.

Despite the immense potential of this strategy, the use of ACT is currently limited because the modified T cells that are transferred back to a person with cancer can be short-lived, and are often unable to overcome a tumour's ability to hinder their function. When naive T cells recognize a disease-causing agent or a tumour cell, they proliferate to form short-lived tumour-killing (that is, cytotoxic) CD8 T cells (also known as effector cells) that kill these infected or malignant cells (Fig. 1). If the infection or the tumour cells are eliminated, most of these CD8 T cells die, but a small population remains in the form of long-lived memory T cells, which are self-renewing and can generate cytotoxic CD8 T cells if the same infection or malignancy is encountered again³.

However, if the infection or tumour cannot be eliminated, the cytotoxic T cells progressively lose their function (a process termed exhaustion). The ideal population of T cells for use in ACT would infiltrate tumours and accumulate in substantial numbers while retaining cytotoxic function and the capacity for self-renewal². Yet the differentiation of T cells into cytotoxic CD8 T cells impairs successful retention of the potential to form long-lived memory cells. This raises the question of whether a strategy can be found to induce both of these beneficial traits in T cells used for ACT. It has been speculated that, in the unforgiving tumour microenvironment, CD8 T cells would need to have a robust metabolism to sustain the nutritional and energetic requirements needed for survival and to retain their antitumour activity⁴.

Wei *et al.* used the CRISPR–Cas9 gene-editing technology to disrupt more than 3,000 genes associated with metabolism in T cells, to test their functions in a mouse model of antitumour ACT. The authors identified more than 200 genes that have a striking ability to affect the persistence and function of the CD8 T cells transferred into tumour-bearing mice. The disruption of many genes had a negative effect on the ability of the cells to persist and thus accumulate in tumours, but the disruption of four genes resulted in a much higher than normal number of T cells infiltrating the tumours.

At the top of this list is the gene that encodes the enzyme REGNASE-1. Its deletion in CD8 T cells caused 2,000 times more of these cells to accumulate in tumours than did CD8 T cells that expressed REGNASE-1. This enzyme binds to and degrades RNA, and influences immune responses^{5–7}, but its role in the antitumour function of CD8 T cells had not been explored. CD8 T cells that lacked REGNASE-1 were better than wild-type CD8 T cells at fighting two types of tumour in mice: an aggressive skin cancer called melanoma and a blood cancer termed acute lymphocytic leukaemia. The REGNASE-1-deficient CD8 T cells proliferated at a similar rate to

Medical research

Antitumour T cells stand the test of time

Miguel Reina-Campos & Ananda W. Goldrath

Enhancing antitumour immune responses has revolutionized cancer treatment, yet some hurdles impede this approach. The discovery of a way to boost the lifespan and function of antitumour immune cells removes a key obstacle. **See p.471**

Immune cells called cytotoxic CD8 T cells can directly kill tumours and are key weapons mobilized in many immunotherapy approaches used in the clinic. However, the cells' activity can be thwarted by the ability of tumours to create harsh microenvironments, recruit immunoregulatory cells and induce inhibitory signals that hamper T-cell function, accumulation and tumour infiltration. On page 471, Wei *et al.*¹ report that depletion of the protein REGNASE-1 extends the survival of antitumour CD8 T cells and enhances their function, enabling the cells to fight cancer more effectively.

The development of anticancer clinical strategies that use immune cells has profoundly improved the treatment of certain malignancies. The delivery of T cells that can specifically target tumours is used in

an approach called adoptive T-cell therapy (ACT), which relies on T cells that have been taken from a person's blood or tumour. These cells are stimulated in the laboratory to cause them to divide and to increase the number of

“This study offers a strong incentive to investigate the use of combinatorial approaches.”

antitumour T cells, and, in some cases, they are modified to enhance their ability to eliminate cancer cells². For example, T cells can be engineered to express a receptor, called a chimaeric antigen receptor (CAR-T), that specifically targets tumours, and such cells are

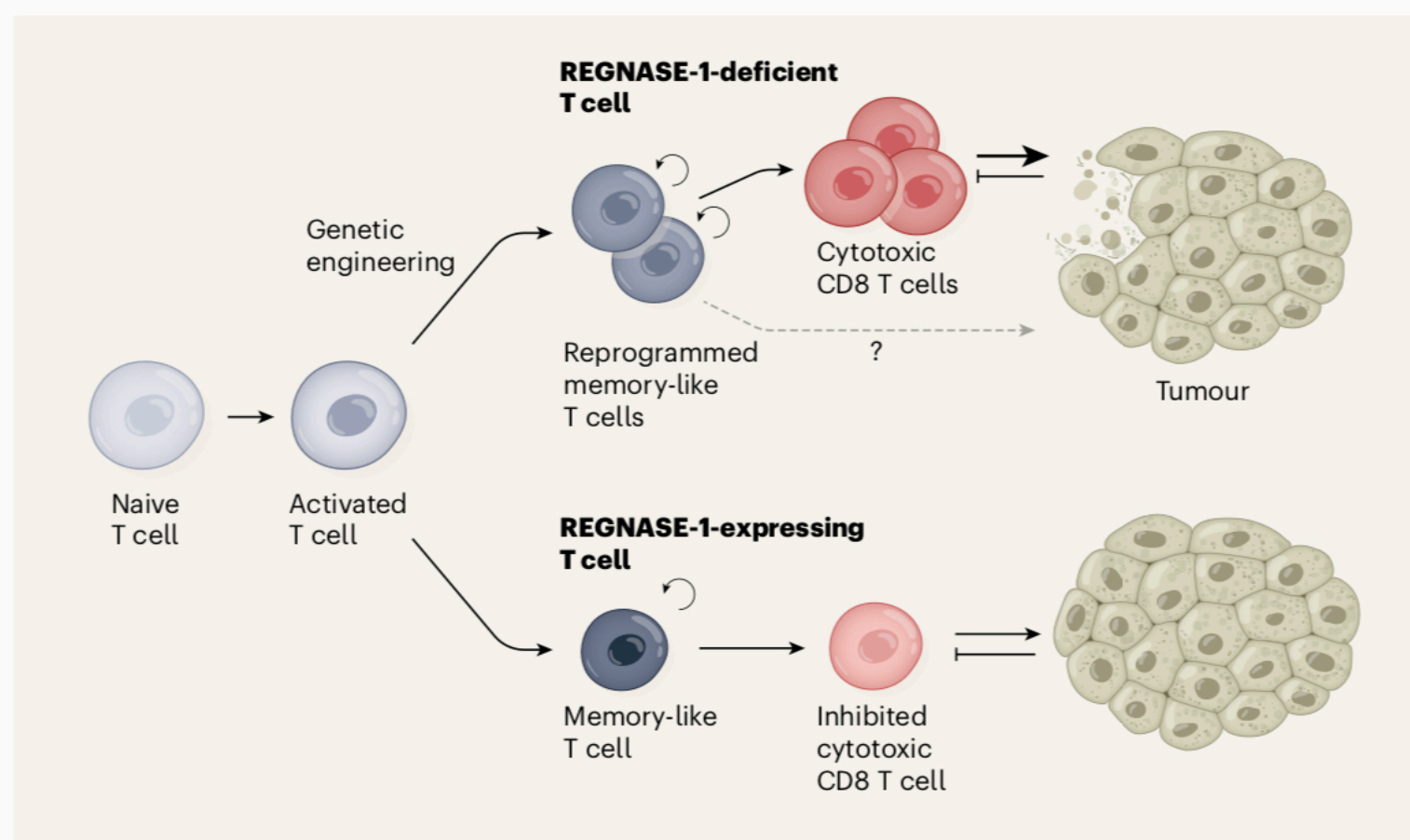


Figure 1 | Boosting T cells to enable a sustained antitumour response. Immunotherapy uses immune cells called T cells to target cancer in the clinic; however, these tumour-killing cytotoxic CD8 T cells are short-lived. Moreover, they can be inhibited by the tumour microenvironment and lose their functional activity because of chronic stimulation if the tumour is not eliminated. Wei *et al.*¹ studied the factors affecting T-cell responses against tumours in mouse models. A T-cell response begins when a naive T cell becomes activated by recognizing its target protein (not shown). Activated T cells can give rise to memory-like T cells that can self renew or generate cytotoxic CD8 T cells. The authors report that if T cells are engineered to lack the protein REGNASE-1, they are reprogrammed to form high numbers of memory-like T cells and have enhanced antitumour activity compared to the case of REGNASE-1-expressing T cells. Therefore, if T cells are REGNASE-1 deficient, they are better equipped to sustain high numbers of cytotoxic CD8 T cells in tumours and produce a prolonged antitumour response. Loss of REGNASE-1 provides the joint benefit of generating T cells that have more tumour-killing activity and memory-like qualities. Whether these reprogrammed memory-like T cells directly contribute to tumour-killing activity is unknown. By contrast, T cells that express REGNASE-1 produce a short-lived antitumour response that fails to overcome tumour inhibition of the immune response.

the wild-type cells, but did not die as rapidly, allowing them to accumulate.

To better understand how REGNASE-1 deficiency resulted in this increased persistence of T cells, the authors analysed gene-expression profiles of REGNASE-1-deficient and wild-type cells. REGNASE-1 deficiency was linked with an increase in a molecular signature characteristic of memory T cells, suggesting the presence of a larger than normal population of long-lived memory-like cells that can give rise to cytotoxic CD8 T cells. REGNASE-1-deficient CD8 T cells showed striking increases in mitochondrial function (mitochondria are organelles that provide a crucial source of cellular energy), including the ability to produce energy and consume oxygen. This is notable because this capacity is often compromised in tumour-fighting T cells⁸. These combined effects of REGNASE-1 deficiency enabled CD8 T cells and CAR-T cells used for ACT to accumulate and remain active over time in the cancers targeted in the mouse models.

To further understand this mechanism, Wei and colleagues used CRISPR–Cas9 to disrupt approximately 20,000 genes in REGNASE-1-deficient CD8 T cells, to pinpoint key downstream genes that mediate the REGNASE-1-dependent cellular reprogramming. The inactivation of the transcription factor BATF, a key regulator

of the differentiation of CD8 T cells⁹, abolished the long lifespan of T cells lacking REGNASE-1 and their high expression of genes associated with mitochondria. The authors found that the combined depletion of REGNASE-1 with that of either of the proteins PTPN2 or SOCS1 had a synergistic effect that increased the persistence, accumulation and antitumour activity of T cells compared with the properties of T cells that were deficient only in REGNASE-1.

Wei and colleagues report that REGNASE-1-deficient CD8 T cells had a higher expression of cytotoxic proteins than did wild-type CD8 T cells in both the memory-like and the cytotoxic CD8 T cells in tumours. Wild-type cells with memory-like properties typically do not kill tumour cells directly¹⁰. It is not clear whether REGNASE-1-deficient memory-like T cells function solely to self-renew and produce the cytotoxic CD8-T-cell population, or whether they can also directly mediate tumour-cell killing, given that their expression of cytotoxic molecules is higher than that of wild-type CD8 T cells.

If they do have a role in tumour-cell killing, how do these REGNASE-1-deficient memory-like CD8 T cells manage to both do this and maintain the population of cytotoxic CD8 T cells? Intriguingly, the authors show that, for increased persistence, the

REGNASE-1-deficient CD8 T cells need to encounter the tumour protein that they recognize. This might explain why the accumulation of antitumour T cells was more pronounced in mouse tumours than in the animals' spleens, which are rich in T cells but are located away from the sites of exposure to the tumour proteins.

It remains to be investigated whether other cues in the tumour microenvironment contribute to boosting the persistence of REGNASE-1-deficient CD8 T cells. To this end, it might be informative to assess the metabolic profile of CD8 T cells in the tumour that have combinatorial depletions of REGNASE-1, PTPN2, SOCS1 and BATF. This could provide insights into the effect of these proteins on the reprogramming of CD8-T-cell metabolism and to what extent this is important for the cells' differentiation and antitumour function. Also, finding the relevant metabolites in this context might provide clues to how these CD8 T cells can be influenced by a nutritionally depleted tumour microenvironment.

Wei and colleagues' study reveals promising leads that might result in advances in ACT-based immunotherapies. It will be worth testing whether engineering CD8 T cells to delete or express low levels of the gene encoding REGNASE-1 would be feasible as part of the manufacturing process for CAR-T cells. Finally, given that the inhibition of PTPN2 in tumour cells sensitizes them to immunotherapy¹¹, this study offers a strong incentive to investigate the use of combinatorial approaches, including REGNASE-1 and PTPN2 inhibitors, as a way to reprogram CD8 T cells to improve current therapies.

Miguel Reina-Campos and Ananda W.

Goldrath are in the Division of Biological Sciences, University of California, San Diego, La Jolla, California 92093, USA.
e-mails: agoldrath@ucsd.edu;
mreinacampos@ucsd.edu

1. Wei, J. *et al.* *Nature* **576**, 471–476 (2019).
2. Lim, W. A. & June, C. H. *Cell* **168**, 724–740 (2017).
3. Henning, A. N., Roychoudhuri, R. & Restifo, N. P. *Nature Rev. Immunol.* **18**, 340–356 (2018).
4. Zhang, L. & Romero, P. *Trends Mol. Med.* **24**, 30–48 (2018).
5. Matsushita, K. *et al.* *Nature* **458**, 1185–1190 (2009).
6. Iwasaki, H. *et al.* *Nature Immunol.* **12**, 1167–1175 (2011).
7. Uehata, T. *et al.* *Cell* **153**, 1036–1049 (2013).
8. Scharping, N. E. *et al.* *Immunity* **45**, 701–703 (2016).
9. Kurachi, M. *et al.* *Nature Immunol.* **15**, 373–383 (2014).
10. Miller, B. C. *et al.* *Nature Immunol.* **20**, 326–336 (2019).
11. Manguso, R. T. *et al.* *Nature* **547**, 413–418 (2017).

A.W.G. declares competing financial interests: see
go.nature.com/2lpjge for details.

This article was published online on 11 December 2019.

Snapshots of science

Neurodegeneration

Selective clearance of mutant huntingtin protein

Huntington's disease is caused by an abnormally long stretch of glutamine amino-acid residues in the huntingtin (HTT) protein. Cells degrade the mutant huntingtin (mHTT) through autophagy – a clearance mechanism that involves engulfment of proteins by a vesicle called the autophagosome. Li *et al.* hypothesized that compounds that bind to both the mutant polyglutamine tract and the protein LC3B, which resides in the autophagosome, would lead to engulfment and enhanced clearance of mHTT. The authors conducted small-molecule screens to identify candidate compounds, and used wild-type HTT in a counter-screen to rule out compounds that bind to the normal version of the protein. They found encouraging evidence that four compounds could produce functional improvements in models of Huntington's disease across three species. This therapeutic strategy might also be useful for other diseases involving expanded polyglutamine tracts.

Huda Y. Zoghbi writing in *Nature* **575**, 57–58 (2019).

Original research: *Nature* **575**, 203–209 (2019).

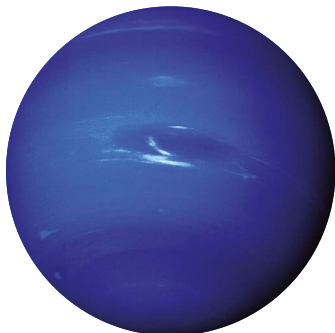
Planetary science

A new moon for Neptune

In 1989, the NASA spacecraft Voyager 2 detected six moons of Neptune that are interior to the orbit of the planet's largest moon, Triton. Showalter *et al.* report the discovery of a seventh inner moon, Hippocamp. Originally designated as S/2004 N1 and Neptune XIV, this moon was found in images taken by NASA's Hubble Space Telescope in 2004–05 and 2009, and then confirmed in further images captured in 2016. Hippocamp is only 34 kilometres wide, which makes it diminutive compared with its larger siblings, and it orbits Neptune (**pictured**) just inside the orbit of Proteus – the planet's second-largest moon. The discovery of Hippocamp is intriguing because of the moon's relationship to Proteus and the role that both objects might have had in the history of Neptune's inner system.

Anne J. Verbiscer writing in *Nature* **566**, 328–329 (2019).

Original research: *Nature* **566**, 350–353 (2019).



Condensed-matter physics

Superconductivity near room temperature

Materials known as superconductors transmit electrical energy with 100% efficiency. They have a wide range of applications, such as magnetic resonance imaging in hospitals. However, these applications have been hampered, largely by the fact that the superconducting state exists only at temperatures well below room temperature (295 kelvin). Drozdov *et al.* report several key results that confirm that, when compressed to pressures of more than one million times Earth's atmospheric pressure, lanthanum hydride compounds, which are rich in hydrogen, become superconducting at 250 K. In the next few years, experiments will probably focus on searching for superconductivity in other pressurized hydrogen-rich materials. Given that only a small fraction of possible hydrogen-rich systems have been subjected to experiments at these tremendous pressures, it seems more likely than ever that the dream of room-temperature superconductivity might be realized in the near future.

James J. Hamlin writing in *Nature* **569**, 491–492 (2019).

Original research: *Nature* **569**, 528–531 (2019).



Fisheries

Micronutrient richness of global fish catches



For the latest News & views published by *Nature*, visit: www.nature.com/research-analysis



Fish are a source of micronutrients that help to prevent nutrient-deficiency diseases. For 43 countries, Hicks *et al.* mapped the relationship between the fish-derived nutrients available from fisheries' catches and the prevalence of such diseases. Their data demonstrate that catches in some developing countries should be enough to meet the key micronutrient needs of their populations. However, in many developing tropical countries, a substantial proportion of local fish catches are either exported or processed locally to generate fishmeal that is then exported and used to feed farmed fish. Many of the local fisheries (**pictured**), which had traditionally supplied the regional markets, now instead supply fishmeal plants. This does not reduce the pressure on wild fish. Moreover, it deprives people on low incomes of previously affordable, nutritious local fish.

Daniel Pauly writing in *Nature* **574**, 41–42 (2019).
Original research: *Nature* **574**, 95–98 (2019).

Genetic engineering

CRISPR tool enables precise genome editing

Tremendous progress has been made in developing gene-editing tools. But a seemingly fundamental limit to the efficiency and precision of gene editing had been reached, owing to the tools' reliance on complex and competing cellular processes. Anzalone *et al.* now describe 'search-and-replace' genome editing, which enables the genome to be altered precisely. In their process, the 'search' part of an RNA guide directs a Cas9 protein to a specific sequence in a DNA target, where it cuts one of the two DNA strands. A reverse transcriptase enzyme then produces DNA complementary to the sequence in the 'replace' part of the guide, and installs it at one of the cut ends, where it takes the place of the original DNA sequence. DNA repair then produces a fully edited duplex. Imperfect edits are almost entirely avoided.

Randall J. Platt writing in *Nature* **576**, 48–49 (2019).
Original research: *Nature* **576**, 149–157 (2019).



Glaciology

Greenland's subglacial methane released

Sediments beneath glaciers and ice sheets harbour carbon reserves that, under certain conditions, can be converted to methane, a potent greenhouse gas. Lamarche-Gagnon *et al.* present direct measurements of dissolved methane in water discharged from a land-terminating glacier of the Greenland Ice Sheet (**pictured**) during the summer. The water was supersaturated with methane, and the amount of methane released to the atmosphere rivals that from other terrestrial rivers. Subglacial sediments can therefore act as a local source of methane, corroborating the results of other studies. Lamarche-Gagnon *et al.* go further by demonstrating that the continuous flux of methane from the Greenland subglacial environment varies with the efficiency of subglacial meltwater drainage. The study provides an example of how our planet's icy domains can interact with the surrounding Earth system in unexpected and potentially important ways.

Lauren C. Andrews writing in *Nature* **565**, 31–32 (2019).
Original research: *Nature* **565**, 73–77 (2019).

Genetics

Fate of a father's mitochondria

The DNA of eukaryotic organisms (such as animals, plants and fungi) is stored in two cellular compartments: in the nucleus and in organelles called mitochondria. A healthy individual's mitochondrial DNA (mtDNA) molecules are mostly identical. However, in people with diseases caused by mtDNA mutations, normal and mutant mtDNA molecules typically coexist in a single cell – a situation termed heteroplasmy. Mitochondrial DNA was thought to derive exclusively from maternal egg cells, with no paternal contribution, but Luo *et al.* challenge this dogma, identifying three families with mtDNA heteroplasmy caused by biparental mitochondrial inheritance. Previous work has shown that mitophagy, the process by which cells 'eat' their own mitochondria, has a role in the selective elimination of paternal mitochondria. These rare instances of paternal mtDNA transmission might therefore be attributed to defective mitochondrial turnover.

Thomas G. McWilliams and **Anu Suomalainen** writing in *Nature* **565**, 296–297 (2019).
Original research: *Proc. Natl Acad. Sci. USA* **115**, 13039–13044 (2018).



Artificial intelligence

Robots on the run

Young animals gallop across fields, climb trees and immediately find their feet with enviable grace after they fall. And like our primate cousins, humans can deploy opposable thumbs and fine motor skills to complete tasks such as effortlessly peeling a clementine or feeling for the correct key in a dark hallway. Although walking and grasping are easy for many living things, robots have been notoriously poor at gauged locomotion and manual dexterity. Even a robot that performs beautifully in simulation will stumble and fall after a few encounters with seemingly minor physical obstacles. Writing in *Science Robotics*, Hwangbo *et al.* report that a data-driven approach to designing robotic software can improve the locomotion skills of robots. They demonstrate their method using the ANYmal robot (pictured) – a medium-dog-sized quadrupedal system.

Hod Lipson writing in *Nature* **568**, 174–175 (2019).
Original research: *Sci. Robot.* **4**, eaau9354 (2019).

Synthesis

Chemical libraries from a double click

A copper-catalysed reaction, known as the CuAAC reaction, is the poster child for click chemistry. A reaction is defined as click chemistry if it is (among other things) operationally simple, high-yielding, applicable to a broad range of compounds, yet exceptionally selective – the chemical groups that undergo the reaction must react only with each other. CuAAC reactions are used in many disciplines, but their applications would be even broader if structurally complex azide compounds (which contain N_3 groups) were more widely available to use as reactants. Meng *et al.* report that fluorosulfonyl azide (FSO_2N_3) reacts with almost any primary amine (compounds that contain NH_2 groups) to achieve a nearly 100% yield of the corresponding azide, and used their reagent to make a library of 1,224 azides. Their reaction meets the speed, breadth and efficiency criteria for click chemistry. Moreover, the prepared azide solutions can be used directly in CuAAC reactions.

Joseph J. Topczewski and **En-Chih Liu** writing in *Nature* **574**, 42–43 (2019).
Original research: *Nature* **574**, 86–89 (2019).

READERS' CHOICE

We asked readers to vote for a News & Views article to be included as part of our round-up of the year. This is the one they chose.

Palaeoanthropology

Unknown human relative found in Asia

Détroit *et al.* report the remarkable discovery of a human relative that will no doubt ignite plenty of scientific debate. This newly identified species was found in the Philippines and named *Homo luzonensis*. Rapidly changing knowledge about hominin evolution in Asia is forcing the re-examination of ideas about early hominin dispersals from Africa to Eurasia. *Homo luzonensis* provides yet more evidence that hints that *Homo erectus* might not have been the only globe-trotting early hominin. The interesting mix of features observed in *H. luzonensis* raises questions about the species' ancestry and its relationships with other human relatives. One thing can be said for certain – our picture of hominin evolution in Asia just got even messier, more complicated and a whole lot more interesting.

Matthew W. Tocheri writing in *Nature* **568**, 176–178 (2019).
Original research: *Nature* **568**, 181–186 (2019).

Why and where an HIV cure is needed and how it might be achieved

<https://doi.org/10.1038/s41586-019-1841-8>

Thumbi Ndung'u^{1,2,3}, Joseph M. McCune⁴ & Steven G. Deeks^{5*}

Received: 23 April 2019

Accepted: 14 November 2019

Published online: 18 December 2019

Despite considerable global investment, only 60% of people who live with HIV currently receive antiretroviral therapy. The sustainability of current programmes remains unknown and key incidence rates are declining only modestly. Given the complexities and expenses associated with lifelong medication, developing an effective curative intervention is now a global priority. Here we review why and where a cure is needed, and how it might be achieved. We argue for expanding these efforts from resource-rich regions to sub-Saharan Africa and elsewhere: for any intervention to have an effect, region-specific biological, therapeutic and implementation issues must be addressed.



**Anniversary
collection:**
[go.nature.com/
nature150](http://go.nature.com/nature150)

Although much effort has been devoted to providing suppressive antiretroviral therapy (ART) to all of those in need, current trajectories suggest that this goal may not be attainable. A safe, effective and durable intervention that completely eliminates the HIV infection (eradication) or that suppresses viraemia in the absence of antiretroviral therapy (remission) (here we refer to both as a 'cure') could serve as an important adjunct in the control of the HIV epidemic (Fig. 1). Although this seems a daunting goal, the scientific motivation is clear: long-term remission if not eradication has been observed in at least two people following transplantation of bone-marrow progenitor cells that lack the viral co-receptor CCR5^{1,2}; and durable remission occurs in approximately 1% of individuals who are infected with HIV (elite controllers) and in 5–10% of those who are treated early in infection and then stop treatment (post-treatment controllers)^{3,4}. Recent advances in animal models suggest that an HIV cure might be induced by some interventions, alone or in combination, such as through the provision of broadly neutralizing antibodies, the generation of an effective antiviral CD8⁺ T cell response or the knockout of CCR5.

Despite some progress, a number of fundamental issues remain (Box 1). There is, for example, no consensus on why a cure is needed. Previous studies have focused on the needs of the individual, but as ART has become safer, more effective and more affordable, it is unclear whether a cure will ever compete with current therapies. In this Review, we argue that the public-health implications of a cure might prove to be at least as important as the benefits for the individual. There is also no consensus on what a cure needs to do. Will a partially effective intervention that provides some with the ability to remain healthy in the absence of therapy for a few years be sufficient, or should we

focus our efforts on complete eradication of the virus? Similarly, will an expensive cure that requires specialized laboratories be useful, or should we seek to develop an approach that is scalable and that could address the main unmet needs globally? Achieving a consensus on these fundamental questions is needed as these decisions will enable the prioritization of competing strategies going forward⁵.

Why and where a cure is needed

There are two main considerations driving the desire to cure HIV: to improve long-term health for individuals who are infected with HIV and to reduce on a community level the transmission of the virus to other individuals. The interests of these two motivating factors are generally well-aligned: for a cure to have an impact on either level, it will need to be safe, effective, durable, scalable and cost-effective. Ideally, it should also protect against reinfection.

The individual perspective

Many individuals living with HIV are able to obtain and adhere to an effective and typically well-tolerated treatment regimen of a single tablet once daily. Assuming such treatment options remain accessible, these individuals have few unmet needs and may not need a cure. Still, not everyone with access to ART does well. Although current regimens are generally safe, they are not benign. Increased risk of cardiovascular, kidney and bone disease has been associated with more-commonly used drugs. Even the recently developed integrase inhibitor class may have long-term health consequences, including weight gain and obesity^{6,7}. Many individuals developed drug-resistant HIV during the early years of the treatment era; complex multidrug regimens are often needed to maintain virus control. As people age and need other medications, polypharmacy has emerged as a concern for nearly all ageing people who have HIV.

Stigma remains a problem for many people living with HIV and is now recognized as a major factor that affects health and well-being⁸. In many communities worldwide, prevalent social constructs make it highly stigmatizing to take ART and/or to attend HIV clinics⁹. Knowledge

¹Africa Health Research Institute, Durban, South Africa. ²HIV Pathogenesis Programme, The Doris Duke Medical Research Institute, University of KwaZulu-Natal, Durban, South Africa. ³Max Planck Institute for Infection Biology, Berlin, Germany. ⁴HIV Frontiers, Global Health Innovative Technology Solutions, Bill & Melinda Gates Foundation, Seattle, WA, USA. ⁵Department of Medicine, University of California San Francisco, San Francisco, CA, USA. *e-mail: Steven.Deeks@ucsf.edu

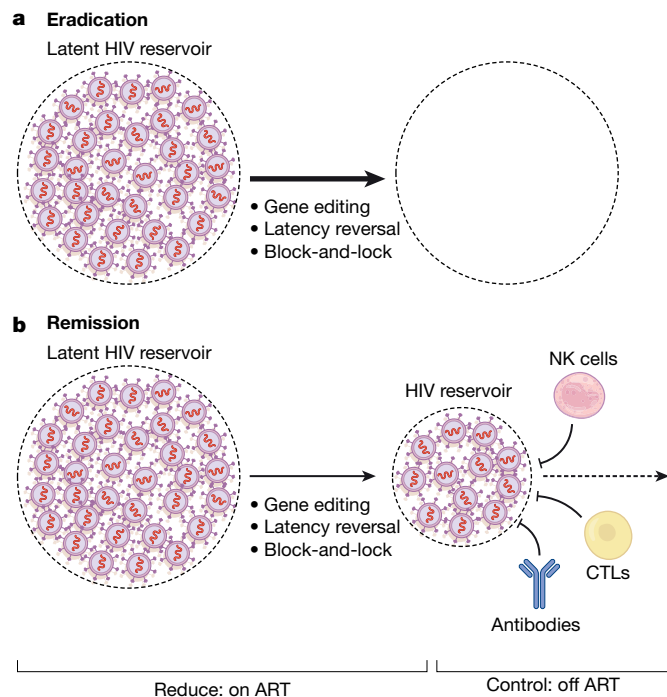


Fig. 1 | Pathways towards a cure. There are two broadly defined pathways for a treatment-free period of virus control. **a**, Eradication. The ideal outcome for a curative intervention would be the complete eradication of all replication-competent virus; gene-editing, latency reversal and block-and-lock approaches are all aimed at reducing the reservoir size and, if fully effective, could lead to complete eradication of the virus within an individual. **b**, Remission. Given observations made in elite and post-treatment controllers, a more plausible strategy may be to reduce the reservoir to more manageable levels while also enhancing immune control. Multiple combination approaches are now being pursued. CTLs, cytotoxic T lymphocytes.

of HIV infection also affects perceptions of social worth and peer relationships¹⁰. It is not yet clear whether an HIV cure can address these challenges, but this would certainly be the goal.

The public-health perspective

Effective control of HIV with ART eliminates the risk of transmission ('undetectable equals untransmissible')^{11,12}, and the ability to test and treat with the aim of reducing transmission is now a universal goal. Currently, approximately half of the infected global population is believed to be on effective ART¹³ and—should this treatment scenario not improve in the future—it is highly unlikely that the epidemic will be contained¹⁴ (Fig. 2). Although nearly all countries have experienced progress in the roll-out of ART, the outcomes have been uneven^{15,16} and the widespread availability of ART has led to only a modest reduction in HIV incidence in some communities with a high prevalence¹⁷.

Antiretroviral treatment programmes in many countries are heavily reliant on international donor partners, raising concerns about the sustainability of these initiatives should the political and economic climate change. Indeed, UNAIDS estimates that US\$26.2 billion in funding (nearly a 30% increase compared to current funding levels) will be required by the year 2020 to achieve their '90–90–90' treatment goals (that is, 90% of people living with HIV should know their status, 90% of those people should be on ART and 90% of those people should have an undetectable viral load)¹⁸. The global commitment to funding is stagnating: whereas the compound annual growth rate for global funding in low- and middle-income countries between 2000 and 2010 was 13%, this number has declined to 1.2% in the last five years.

Theoretically, a short-term, affordable and effective intervention that results in sustained periods of virus control in the absence of any

therapy will reduce the overall strain on public-health systems, freeing up resources for other healthcare imperatives including other aspects of HIV prevention and care. Ideally, such a regimen will need to be as safe and effective as ART—a bar that might not be possible to achieve, particularly in adherent populations in which treatment response rates now approach 100%. This raises questions of how such a regimen will obtain regulatory approval and how it will be implemented. Discussions among key stakeholders will be needed to define in which situations a regimen that is less effective than ART will be effective and how such an approach will achieve regulatory approval.

On a global level, approximately 20% of individuals with HIV do not know their status¹³ (Fig. 2). Multiple barriers to getting tested exist, including stigma and concerns about being diagnosed with an incurable disease. The availability of an effective cure might boost HIV control programmes by encouraging disenfranchised individuals with HIV to proactively seek testing and treatment for HIV, as has been documented for syphilis¹⁹.

Why ART is not curative

Many of the curative interventions that are currently being explored address virological and immunological factors that limit the ability of ART itself to cure HIV infection, as described below.

Latency

As a retrovirus, the HIV genome fully integrates into the host genome and persists for the lifetime of the infected cell^{20–22}. The vast majority of these integrated genomes appear to be transcriptionally silent (that is, latent)²³. This state of latency—in which viral proteins are not produced—enables infected cells to escape immune recognition and clearance. Latency is maintained by multiple mechanisms, including expression of unique and complex transcriptional pathways that may prevent reactivation²⁴, the upregulation of anti-apoptotic genes²⁵ and blocks to various post-initiation transcriptional pathways (such as elongation, polyadenylation and multiple splicing)²⁶.

Reservoir dynamics

Approximately 0.01–1% of circulating CD4⁺ T cells contain an integrated genome, only a small proportion (<5–10%) of which are fully intact²⁷, and only a small proportion of intact genomes may be readily inducible and able to support virus replication post-ART^{27,28}.

There is intense interest in further defining the cell populations that are more likely to be infected. CD32 was reported to be a putative biomarker²⁹, but subsequent studies have failed to confirm this finding. PD-1 has been associated with the reservoir in several studies^{30–32}. Other cell populations enriched for HIV include those that express other checkpoint receptors^{31,33,34}, markers of activation or proliferation³⁵, members of the tumour necrosis factor family^{25,36} and markers of cell adhesion and migration^{37,38}. The virus may also be enriched in the more differentiated effector memory cell population^{35,37}, such as the T helper 1 and 17 subsets^{39,40}. As these associations are modest and highly variable, none will prove useful as biomarkers or as targets for host-directed immunotherapies. These findings, however, provide insights into how HIV establishes latency.

The fate of latently infected cells is largely dictated by the physiological pathways of T cell homeostasis^{24,30,41–43}; infected cells undergo clonal expansion and are then maintained by those same factors that control the size and diversity of the memory T cell pool⁴¹. During long-term ART, individual clonal populations wax and wane as the entire reservoir becomes increasingly clonal in nature⁴⁴. Some integration events can disrupt the regulation of cell growth, leading to massive expansions^{42,43}. The local chromatin environment in which the virus integrates is also important: genomes integrated in largely silenced regions are less likely to be inducible^{45,46}; indeed, some may be permanently silenced. These silenced genomes may be selected for and enriched during long-term ART.

Box 1

Unresolved questions

The road to an HIV cure will be long and unpredictable. There are, however, several important issues that can be acknowledged and addressed.

The nature of a curative intervention. Complete eradication of the virus from an individual is the ideal outcome of a cure intervention; however, short of aggressive interventions (for example, haematopoietic stem-cell transplantation), this may not be possible, at least with current technologies. Short of such a cure, the induction of a durable immune response (leading to sustained ART-free remission) is appealing. Such a response must be primed, expanded and topologically disposed to prevent viral spread as soon as it begins. Ideally, it would prevent rebound of the virus from endogenous sources and block viraemia upon re-exposure. It is likely that combination approaches will be required and, given the distribution of the epidemic around the world, curative interventions of this type must be safe, effective and scalable to be truly effective. It is hoped that future advances in gene targeting and editing *in vivo* may eventually lead to a more definitive and scalable cure.

The clinical evaluation of candidate cures. At present, treatment interruption is the only way to discern the effect of most interventions. Although strategies have been developed to address associated risks to the treated individual, such interruptions also pose substantial risks to sexual partners, and there is no consensus about the mitigation of these risks. A high priority now is to identify circulating non-viral biomarkers that can predict the rebound of infectious virus. Ideally, such biomarkers would ultimately form the substrate of diagnostic tests that could be used by individuals to detect the failure of a candidate cure—that is, viral rebound—should it occur.

Emerging data suggest that the intact genome may decay more rapidly than the defective genomes^{47,48}, although some studies have argued the opposite⁴⁹. It has also been suggested that the inducibility of intact genomes declines over time as the reservoir increasingly enters a relative state of deep latency^{45,47}; if confirmed, this would suggest that virus populations during long-term ART will be less likely to initiate rounds of virus replication after the cessation of therapy.

HIV replication

HIV may also persist in the face of suppressive ART owing to low but detectable levels of virus replication in lymphoid microenvironments⁵⁰, although the fact that the virus does not evolve over time argues that ART can be fully suppressive⁵¹. Even if this is the case, ART can only prevent cells from becoming newly infected; cells that are already infected must be eliminated by other means.

Host clearance mechanisms

The capacity of the immune system to clear infected cells is also likely to be an important factor that contributes to the persistence of HIV⁵². During acute infection, the virus evolves rapidly to escape the immune system and these mutants are retained in the reservoir⁵³. The numbers of HIV-specific CD8⁺ T cells during ART are low (as expected given low antigen levels), often dysfunctional (as defined by expression of PD-1 and other markers) and often target HIV variants that have already escaped immune recognition^{54–57}. The administration of ART during acute HIV infection may prevent many of these abnormalities from emerging, but cells are often not sufficiently primed to be effective^{58–60}. Many of the immunotherapies that are under development for an HIV cure seek to reverse these defects.

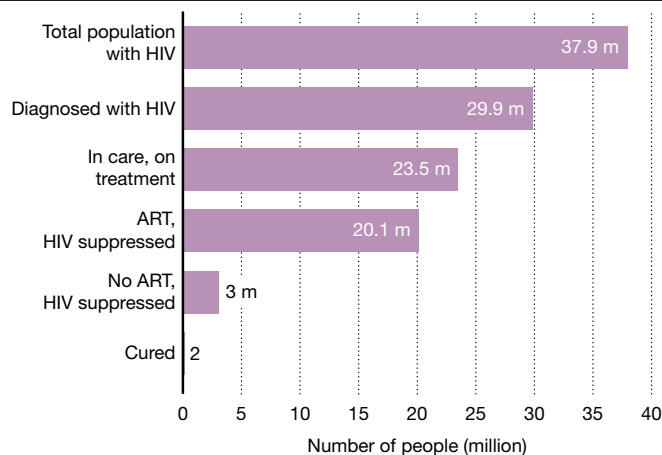


Fig. 2 | The cascade of treatment and control. According to the most recent UNAIDS estimates, of the 37.9 million people living with HIV, only about 79% have been diagnosed and only about 53% are on effective therapy. It is estimated that about 1% are doing well in the absence of therapy ('elite' control and rarely 'post-treatment control'). Only two individuals are believed to have been cured.

How to define a cure

The ideal outcome for any curative intervention would be the complete eradication of all replication-competent HIV particles—that is, a sterilizing cure (Fig. 1). As it will not be possible to prove that all infected cells that have replication-competent HIV have been eliminated⁶¹, a potentially cured individual will never really know whether any virus particles persist that can restart a systemic infection. Indeed, in several cases of very early ART treatment or allogeneic bone marrow transplants, the virus was no longer detectable yet rebounded within weeks to months after ART interruption^{62–65}. As an infected cell can initiate acute viraemia at any given time, new point-of-care or preferably at-home tests that can reliably determine whether the virus has rebounded may prove to be necessary for any strategy that is less than 100% effective⁶⁶. This inability to predict whether and when a rare infected cell will initiate an episode of acute viraemia argues for interventions that confer durable host-mediated control of a residual reservoir.

Most strategies that are currently being pursued seek to durably induce a state in which the virus is maintained at such low levels that it can no longer cause disease or be transmitted to others (Fig. 1). Operationally, such a functional cure or state of durable virus remission will probably be defined as the sustained maintenance of a viral load that is less than the level of quantification with standard assays (for example, 40–50 copies of RNA per ml of plasma for some yet-to-be-defined period). Of note, remission strategies that achieve a low but persistently detectable level of viraemia may provide important benefits for individuals who are unable to adhere to standard regimens, including those with multidrug resistance or disenfranchised individuals who are unable to access or adhere to therapy.

How to measure a cure

A major barrier to the development of curative interventions is the lack of a well-validated surrogate biomarker for the total reservoir of replication-competent virus in the body. This is especially crippling for 'go/no-go' decisions at the proof-of-concept stage and a key impediment to the initiation of investment in this new therapeutic area by biotechnological or pharmaceutical companies.

Although HIV DNA can be readily measured in the CD4⁺ T cell compartment of ART-suppressed individuals, most of it (>95%) is defective and unable to contribute to new rounds of infection within the host or be transmitted to others²⁷. Assays that directly quantify or estimate the

Box 2

TPP of a potential cure

Efforts are underway to identify those characteristics of a curative intervention that will be necessary for the intervention to be effective. It is expected that the first generation of cures will be expensive, require access to advanced technologies and have limited scalability. For a cure to truly alter the course of the global epidemic, a number of factors will need to be considered, all of which contribute to the target product profile (TPP). The minimal and optimal characteristics of a cure TPP should be discussed among all stakeholders as early in drug development as possible. Here we discuss the factors and associated characteristics that should be considered for an intervention to be effective.

- **Clinical effect.** The desired outcome (such as remission or the elimination of viral reservoir) and efficacy threshold should be well-defined.
- **Indication.** The target populations to be included (for example, adults and children with HIV viral load < 100,000).
- **Re-exposure and reuse.** The number and frequency of doses required for the cure to be effective and the requirements for pretreatment.
- **Dosing and administration.** In addition to the number and frequency of doses required for the cure to be effective and the requirements for pretreatment, the treatment route (such as infusion or oral administration) should be considered.
- **Storage and handling.** The supply chain (for example, cold storage), formulation (such as the need for compounding) and shelf life need to be assessed.
- **Follow-up.** Clinical monitoring, biomarker testing (such as viral load) and confirmatory testing (test of latent reservoir) need to be carried out.
- **Contraindications.** Comorbidities (such as chronic kidney disease), concurrent therapies, social factors (that is, adherence to the treatment regime) and other conditions (such as pregnancy) need to be taken into account.
- **Safety and toxicity.** Side effects (such as nausea, diarrhoea and mood disturbances) and toxicity (to the renal or hepatic system, for example) need to be assessed.
- **Cost.** The goal price in target markets needs to be such that this intervention will reach those people with the highest need.

circulating reservoir of replication-competent virus are under development⁶⁷, but the degree to which these approaches quantify viruses that can initiate new rounds of infection (the rebound-competent viral reservoir) has yet to be determined. Additionally, it is not clear whether and to what extent any measurement using blood will be able to quantify the level and disposition of viral reservoirs that are contained within tissues, where the vast preponderance of the replication-competent pool resides⁶⁸. Novel approaches that quantify the virus within tissues directly (for example, using biopsies or imaging using labelled anti-HIV antibodies) or indirectly (for example, by assessing the host response to the virus by quantifying antibody titres) are in development^{69,70}.

Short of an unexpected breakthrough across the above barriers, the only way to determine the effectiveness of a putative curative intervention is to interrupt treatment with ART and then measure either the time-to-rebound (test of cure) or the level of set-point viraemia after rebound. The former approach is relatively safe if the viral load is measured frequently (usually weekly) and ART is resumed immediately upon the detection of rebound^{62,63,71}. Defining the post-interruption set point is more informative but far more complicated. People who

are destined to control their virus in the absence of therapy may first experience a rapid but transient burst in viraemia^{3,72}, which can cause harm to the immune system of the individual and poses risks to sexual partners⁷³. Strategies that aim to reduce the risk to participants and their sexual partners have been developed⁷⁴.

What a cure needs to achieve

Cure research remains focused in academic medical centres located in low HIV burden, resource-rich countries. As such, there has been limited public discussion on the practicalities of product development, particularly as they relate to sub-Saharan Africa. Failure early on to define a target product profile (TPP) risks developing a strategy that fails to be effective. In principle, a TPP should provide a broad outline of the minimal attributes of an effective therapy, guiding drug development while engaging all stakeholders, including the community, the regulators, the funders, the implementers and industry (Box 2).

An unresolved question is whether a cure will need to compete with and potentially replace ART. If this is the case, it will have to be as safe and effective as standard ART regimens, including emerging long-acting formulations—a high bar that is unlikely to be reached by currently available interventions (for example, CCR5Δ32 bone-marrow transplantation).

As modern ART regimens are remarkably safe and well-tolerated, any intervention associated with serious toxicities will probably be unacceptable to most stakeholders. With regard to efficacy, short-term failures in which a cure is administered and then quickly found to have failed could easily be managed by continuing or resuming ART. Long-term failures in which the virus rebounds after an unpredictable period of time would potentially be disastrous. In this setting, a burst of viraemia may go unnoticed for months, leaving the person and his or her sexual partner(s) at risk. Modern cure studies generally have indefinite periods of close monitoring to avoid this scenario⁶³. Inexpensive at-home diagnostics that are able to rapidly detect rebound might be needed when a curative strategy is implemented.

Other factors are probably also important. If a cure leaves an individual susceptible to reinfection, then at least one model suggests it will rarely ever be effective from a public-health perspective, given that those who acquire HIV once will presumably have sustained exposures and risks for acquiring the infection a second time¹⁴. The 'Berlin patient', for example (one of the two known cases in whom HIV is potentially cured), has disclosed publicly that he is now taking pre-exposure prophylaxis (PrEP); in other words, he has switched from a three-drug to a two-drug regimen and is not living in a state of ART-free viral suppression.

A critical consideration for resource-limited settings will be ease of administration. Healthcare systems in most low- and middle-income countries are already strained by the existing disease burden and will not be able to cope with cure strategies that require the administration of multiple products, inpatient evaluations, a specialized infrastructure or extensive cold chain management.

Finally, a curative intervention will only have a global influence if it is cost-effective. In the context of the epidemic in Zimbabwe, for instance, an accessible intervention would only make sense financially as a replacement for ART if it costs less than US\$1,400 and enables ART interruption in 95% of treated individuals with a rate of viral rebound of approximately 5% per year⁷⁵. For South African people with access to effective therapy, only a safe and highly effective curative intervention would be cost-effective⁷⁶.

How to achieve a cure

There are five broadly defined approaches to achieving an eradication cure or remission: early ART, genetic modifications, 'shock-and-kill', 'block-and-lock' and immunotherapy.

Early ART

Starting therapy immediately after the diagnosis of infection with HIV preserves immune function, limits virus diversification, minimizes HIV-related complications in the treated individual and prevents transmission to others by lowering viraemia^{59,77}. Although early ART may also prove to be curative (as suggested in a study of non-human primates)⁷⁸, experience to date suggests that this will not often be observed in adults who are infected with HIV. Thus, in a carefully performed study of ten adults in Thailand who started ART in the first few weeks after infection, the virus rebounded almost immediately after ART was interrupted⁶². Similarly, an individual who started pre-exposure prophylaxis immediately (probably within a day or two) after a recent infection nevertheless experienced viral rebound once ART was discontinued⁶³.

Although immediate ART will probably never cure an adult, it might under some circumstances provide benefits to infants. In the well-known case of the 'Mississippi baby', for instance, an infant was infected in utero, born with a viral load of approximately 20,000 copies of RNA per ml, started on ART within 30 h of her birth, stopped therapy 18 months later and then experienced 30 months of virus-free remission before the occurrence of rebound⁶⁵. This case raises the possibility that treatment in infants might prove curative if started earlier or maintained longer.

Even if not curative, early ART appears to reshape the association between the virus and the immune system, allowing some people to eventually control their virus in the absence of treatment. Approximately 5–10% of those treated early during the course of disease develop a state of ART-free viral remission^{3,4}. These so-called post-treatment controllers generally start ART in the first few weeks to months of their infection, remain on ART for several years and then stop therapy for various reasons⁴. Less commonly, post-treatment control can occur in perinatally infected children who begin ART treatment soon after birth^{79,80}.

Genetic modifications and cell therapy

There are two known cases in whom HIV is potentially cured: the Berlin patient¹ and the 'London patient'². In each case, an HIV-seropositive patient with cancer received fully ablative chemotherapy and an allogeneic stem-cell transplant from donors who naturally lacked CCR5 (CCR5Δ32/Δ32), a co-receptor used by many circulating virus strains to enter CD4⁺ T cells. Once 100% chimaerism was achieved, all available CD4⁺ T cells were resistant to new infections and the virus was unable to spread to uninfected cells. Although associated with multiple toxicities and high cost, these cases have provided proof-of-concept evidence that genetic and/or cellular modifications of the affected host might lead to a state of durable viral remission.

Given these cases, the question arises of whether there are safer and more scalable ways to achieve the same result. The two transplant cases and limited studies in animal models^{81,82} argue that, if CCR5 can be safely disrupted in T cells (including key progenitor cells), a cure may be achievable. As there are a number of approaches to genetically edit or suppress the function of CCR5⁸³, a major hurdle is the manner of delivery, which ideally would be done in vivo and not require ex vivo manipulation and transplantation. Given the strong interest in developing in vivo gene-editing strategies in other areas of medicine⁸⁴, it seems likely that a safe and scalable delivery approach will eventually be able to target gene-modifying systems to appropriate, long-lived cell populations in vivo, precisely and safely disrupting or inhibiting the function of CCR5 and leading to an effective cure. CCR5 may be an ideal candidate for such a technology as its disruption would probably prove to be safe and effective. There may be a threshold below which the frequency of susceptible cells is insufficient to support systemic infection⁸⁵, arguing that even incomplete editing will be curative.

Although editing CCR5 is attractive, the limitations of this approach are not trivial. Many individuals have viruses that utilize CXCR4 for

cell entry. Indeed, one recipient of an allogeneic stem-cell transplant from a donor with homozygous deletion of CCR5 stopped therapy and exhibited rapid rebound of a pre-existing CXCR4-utilizing variant⁸⁶. Furthermore, although individuals who are homozygous for the CCR5Δ32 allele generally do well, it remains unknown whether disruption of this pathway by gene modification will be safe.

Gene-editing approaches that target and excise the integrated provirus would also be potentially curative⁸⁷. Nucleases that are specific to conserved areas of the provirus have been developed and validated in vitro, although selection for virus populations that are resistant to editing can occur. Combination approaches that target multiple sites in the virus may be needed, just as a combination of antiretroviral drugs is necessary to prevent HIV resistance⁸⁸. Delivering a universally effective combination of gene-editing enzymes to all cells in the body that are infected with replication-competent viruses will be challenging but necessary for this strategy to work.

Alternatively, novel genetic modifications that lead to an eradicated cure or remission might be delivered. Theoretically, if a cell can be re-engineered to produce an effective antiviral response indefinitely, then a state of durable control will be achieved. In a proof-of-concept study, SHIV-infected rhesus monkeys were infected with adeno-associated virus vectors that encoded three potent neutralizing antibodies against HIV; in one animal, high levels of the antibodies were durably produced, resulting in sustained control of the virus⁸⁹. Efforts to repeat this in humans have been slowed by the development of anti-drug antibodies to the vector⁹⁰.

Studies of individuals who naturally control HIV in the absence of therapy (elite controllers) have demonstrated that functional HIV-specific CD8⁺ T cells that target vulnerable regions of the virus can control the virus for years, suggesting that this mechanism may be harnessed for a functional cure. Although early attempts to re-engineer CD8⁺ T cells to express a chimeric antigen receptor (CAR) that targets HIV failed⁹¹, more mature versions of this technology are now being used for the management of B cell malignancies⁹² and actively repurposed for studies researching an HIV cure^{93–95}. Barriers that will need to be overcome include limited sensitivity (the density of virus proteins on the surface may be too low to be readily recognized), limited persistence of the CAR-T cells in vivo, escape mutations and safety.

Latency reversal

With the hope that the induction of viral proteins could lead to immune-mediated recognition and destruction of infected cells, attempts have been made to reverse latency (called shock-and-kill or 'kick-and-kill'). Multiple first-generation latency-reversing agents have been studied in the clinic, and several have been shown to stimulate the transcription of HIV mRNA^{96–98} and perhaps even the production of viral particles^{99,100}. A second generation of approaches has focused instead on non-specific inducers of T cell activation, for example, Toll-like receptor (TLR) agonists^{101,102}. Given the complexity and heterogeneity of the reservoir⁴⁶, it seems likely that combinations of approaches may be needed^{103–105}.

A fundamental problem with latency reversal is that, in the absence of the concomitant generation of other factors (for example, an immune response that can suppress viraemia), it may be necessary to eliminate most if not all of the reservoir. Although one model predicts that a 10,000-fold reduction in the reservoir will be needed to prevent virus rebound¹⁰⁶, recent clinical data suggest that even with exceedingly small and undetectable reservoirs the virus will rebound within months⁶³. As complete or near-complete elimination may prove to be impossible with latency-reversing agents alone, these approaches are now being proposed as a way to reduce the size of the reservoir to a more manageable level, possibly enhancing the effectiveness of other (for example, immune-based) strategies (Figs. 1, 3). This approach is analogous to tumour debulking in cancer therapeutics.

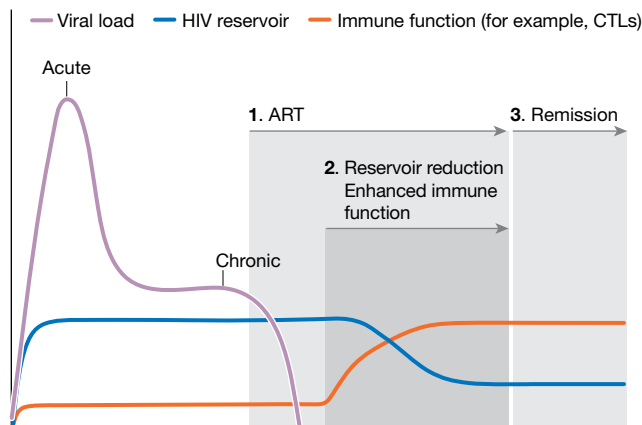


Fig. 3 | HIV remission pathway. Many strategies that are currently in development are aimed at achieving durable control, rather than complete eradication. All interventions require individuals to first control HIV with ART. Interventions that are aimed at reducing the reservoir size (shock-and-kill, block-and-lock) are then initiated, followed by combination approaches to enhance immune control. ART is eventually interrupted, followed by an expected transient period of transient viraemia and eventually control of HIV. Post-ART monitoring of viral load may prove to be the most challenging and expensive component of this strategy.

Deep latency

In studies carried out *ex vivo*, it has been found to be difficult to induce some proviruses out of latency^{27,28}, suggesting that inducing a state of deep latency might contribute to a cure. During long-term ART, intact genomes may become enriched in non-genic chromosomal regions that are less likely to support efficient transcription of the viral genome⁴⁵. Viral genomes that are therapeutically placed into such a state of deep latency would be effectively eliminated as a source of virus recrudescence. For example, didehydro-cortistatin A¹⁰⁷ (a Tat inhibitor) caused a modest delay in virus rebound post-ART in a humanized mouse model¹⁰⁸. As with the latency reversing approach, it seems likely that this approach (generally referred to as block-and-lock) might work best to reduce the effective reservoir size, thus enhancing the efficacy of other approaches.

Immunotherapy

Propelled by revolutionary advances in the field of oncology, multiple ongoing or planned clinical trials are testing various immunotherapies in people with HIV¹⁰⁹. Most of these studies have targeted CD8⁺ T cells and multiple factors probably account for the fact that all have failed. The size of the reservoir in some people may be too large¹¹⁰. Many of the initial CD8⁺ T cell responses target highly variable regions of the virus, resulting in rapid selection of escape mutants⁵³. Sustained exposure to the antigen drives T cells towards a dysfunctional state marked by upregulation of PD-1 and other regulatory pathways. These immunodominant and dysfunctional responses that target escaped epitopes are typically the first to be expanded therapeutically with vaccines and other approaches¹¹¹. HIV infection induces a potent inflammatory response that is reduced but not eliminated by ART; this inflammatory state induces counter-regulatory responses that prevent T cells and other effectors from working optimally. Finally, some of the reservoir resides in the B cell follicle³², an anatomical space from which CD8⁺ T cells, natural killer (NK) cells, and other effector responses are typically excluded. For an immunotherapeutic regimen to work, it may be necessary to breach these barriers (Figs. 1, 3).

There are, however, examples of success, particularly in the non-human primate model. A re-engineered replication-deficient CMV vector with simian immunodeficiency virus (SIV) inserts induced an unusual (MHC class II- or MHC-E-restricted) T cell response of unprecedented breadth¹¹². Although the vaccine failed to prevent SIV acquisition, it

did lead to control and eventually the elimination of the infection in approximately 50% of exposed animals¹¹³. The administration of broadly neutralizing antibodies to infant macaques either cleared or prevented the establishment of a latent reservoir¹¹⁴. A therapeutic vaccine and a vaccine adjuvant (a TLR7 agonist) induced post-ART remission mediated by CD8⁺ T cells in a subset of SIV-infected monkeys¹¹⁵. In ART-treated macaques, a TLR7 agonist alone was shown to induce the expression of SIV RNA, a reduction of viral DNA, the activation of innate and adaptive immune cells and viral control for more than two years¹⁰¹. A broadly neutralizing antibody that targets the viral envelope combined with the same TLR7 agonist either eliminated the reservoir or generated a post-ART state of remission¹¹⁶. Finally, therapies using broadly neutralizing antibodies during acute SHIV infection induced subsequent remission¹¹⁷, possibly through a 'vaccinal effect', as described below. These approaches are now being translated into clinical studies.

Arguably, broadly neutralizing antibodies have been the linchpin in most of the successful immunotherapeutic studies. When used in combination, broadly neutralizing antibodies have a potent antiviral effect that is comparable to that of ART¹¹⁸. In contrast to ART, however, broadly neutralizing antibodies can theoretically clear reservoir cells by stimulating antibody-dependent cellular toxicity and other Fc-receptor-dependent cytotoxic effects. Antibody-antigen immune complexes can also activate Fc receptors on antigen-presenting cells and induce a sustained T cell effect (a vaccinal effect). A combination of two antibodies may have induced long-term remission in two individuals through such an effect¹¹⁸. A newer generation of antibodies that bind to two or three sites on the viral envelope (bi-specific and tri-specific antibodies) has been shown to be effective in a macaque model¹¹⁹ and the use of such antibodies is moving towards clinical applications. Dual-specific antibodies that bind to HIV and the T cell surface receptor CD3 have the potential to stimulate T cells to target and eliminate HIV-infected target cells¹²⁰.

Therapeutic vaccines—fuelled by innovations in adjuvants, immunogen design and delivery modalities such as novel DNA and RNA systems, viral vectors and *ex vivo* loading of dendritic cells—are showing promise in the treatment of cancer and chronic infections, including HIV¹²¹. Most of these vaccines are designed to target conserved or vulnerable regions of the HIV virus¹²².

Borrowing heavily from the oncology field, ongoing studies are testing agents that directly stimulate or redirect immune responses (for example, the cytokine IL-15 and TLR agonists), or that reverse the dysfunction of T cells that is induced by chronic exposure to HIV (for example, inhibitors of PD-1 and CTLA-4). These agents are inherently activating and hence might have the added benefit of reversing latency. PD-1 inhibitors, for example, induce virus production *ex vivo*¹²³ and may enhance HIV-specific T cell responses *in vivo*¹²⁴, thus resulting in a combined shock-and-kill effect.

Non-specific therapies that reverse the chronic inflammatory state of treated HIV might induce a more-effective immune response. Inhibitors of IL-10, IFN α , TGF β , mTOR, IL-1 β , JAK-STAT, and other pathways are being tested in animals and/or people. Although most of these approaches are being pursued as potential strategies to reduce the burden of inflammation-associated co-morbidities in treated HIV disease, some might plausibly enhance the ability of the adaptive immune system to reduce or control the reservoir.

To facilitate the entry of CD8⁺ T cells and NK cells into the B cell follicle, agents that deplete B cells (rituximab) or that enhance the migration of CD8⁺ T cells (IL-15) are being tested as adjuncts for immunotherapies in animal studies¹²⁵. By slowly reducing the inflammatory state that maintains inflamed follicles and the production or replication of the virus^{32,126}, long-term ART might prove to be the most effective way to deal with this potential barrier.

Combination approaches

Most studies of curative interventions that are now in the clinic are designed to test safety and to provide proof-of-concept that a relevant

Box 3

Unresolved questions regarding cure studies in sub-Saharan Africa

Arguably, an effective strategy for a HIV cure is most urgently needed in sub-Saharan Africa, where the prevalence and incidence of HIV are highest. However, most studies that have tried to find a HIV cure have been conducted in low-burden, resource-rich countries. Unless this research imbalance is corrected, there is a potential to develop cure strategies that may not work in all populations, or that may prove difficult to scale and implement in those communities in which the need is greatest. Some of the critical questions that need to be addressed are:

1. What do those individuals living with HIV in Africa want from a cure?
2. Who are the target populations that might best benefit from a cure?
3. What are the characteristics of an intervention that lead to an eradicated cure or remission that would be best suited to these target populations?
4. What will a cure need to cost to be accessible and affordable?
5. What kind of interventions can be practically given in non-urban healthcare clinics?
6. Will the HIV subtype have an effect on the effectiveness of current cure interventions?
7. Will therapies developed for subtype B (for example, vaccines and antibodies) need to be redesigned for other subtypes?
8. Will persistent inflammation be a concern? Will common prevalent co-infections (for example, with tuberculosis, malaria and helminthic worms) affect or preclude the use of immunotherapy?
9. Are the replication-competent reservoirs that are found among African people qualitatively and quantitatively similar to those studied elsewhere (for example, in the United States and Europe)? In particular, might differential aspects of age, gender, race and concomitant co-infections have an effect?
10. Will there be a practical, scalable, affordable and acceptable method available for detecting failure (that is, rebound viraemia), should that occur?

pathway is effectively targeted, and it is not likely that any of these approaches will effect an eradicated cure or remission when used alone. There is accordingly an evolving perspective that the field should more rapidly advance combination therapies that are designed to achieve a complete cure or remission^{127,128}. A popular combination approach is to reduce the reservoir to levels that a therapeutically enhanced immune system can control ('reduce and control'; Figs. 1, 3). Clinical trials using analytical treatment interruptions are increasingly being used to study these combination approaches⁷⁴.

What is needed to achieve a scalable cure

Research on a cure in those areas of the world where the need is greatest is essentially non-existent. From a biological perspective, there are many virological, immunological and socioeconomic factors that are unique to Africa and elsewhere that might conceivably affect the impact of a cure intervention^{129,130} (Box 3). The HIV-infected populations in resource-rich areas are older and predominantly male, whereas in Africa

they are younger and predominantly female. Recent evidence suggests that ART will have differential toxicities in these populations^{6,7} and that sex influences reservoir activity and the immune response¹³¹. From an implementation perspective, failure to develop an intervention that is effective, affordable and easy to administer risks limiting the ultimate impact of the strategy, and failure to engage communities early on in the research process could result in delays in implementation that cost lives when a product becomes available.

The lack of adequate clinical and laboratory infrastructure is a major barrier to expanding the cure agenda to Africa. Currently, most HIV clinical research sites in Africa perform observational cohort studies or late-stage clinical trials of ART or vaccines, with little involvement in early-stage clinical research. If resource-poor, high-burden areas such as Africa are to play a more prominent part in this endeavour, the laboratory and clinical infrastructure needed to support this kind of research will require substantial improvement. Considering the previous experience with antiretroviral drug research in Africa, these are not insurmountable problems.

Outlook

Although HIV treatment with ART has improved public health worldwide and resulted in a decrease in the incidence of HIV in most countries, lifelong treatment poses financial, logistical, health and psychosocial challenges. A safe, effective, scalable and cost-effective intervention that is fully curative or that allows for a sustained period of virus control in the absence of any therapy would provide a powerful adjunct for the eventual control of the epidemic. Although research towards developing such an intervention remains in its infancy, recent observations provide insight, suggest progress and prompt continued interest in experimental medicine studies and small early-phase clinical trials of promising strategies. Ultimately, for an intervention that leads to an eradicated cure or remission of HIV disease to reshape the course of the epidemic, these studies will need to be conducted in communities that are most likely to benefit from such research.

1. Hütter, G. et al. Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. *N. Engl. J. Med.* **360**, 692–698 (2009).
2. Gupta, R. K. et al. HIV-1 remission following CCR5Δ32/Δ32 haematopoietic stem-cell transplantation. *Nature* **568**, 244–248 (2019).
Two HIV-infected adults¹² with haematological malignancies were apparently cured of HIV after an effective stem-cell transplant from an allogeneic donor whose T cells lacked CCR5, a key co-receptor for virus entry.
3. Namazi, G. et al. The control of HIV after antiretroviral medication pause (CHAMP) study: posttreatment controllers identified from 14 clinical studies. *J. Infect. Dis.* **218**, 1954–1963 (2018).
4. Sáez-Cirión, A. et al. Post-treatment HIV-1 controllers with a long-term virological remission after the interruption of early initiated antiretroviral therapy ANRS VISCONTI Study. *PLoS Pathog.* **9**, e1003211 (2013).
A subset of HIV-infected adults who start ART early and remain on therapy for a sustained period are able to effectively control HIV replication after treatment interruption; although the mechanism(s) at play remain unclear, these 'post-treatment controllers' provide strong evidence that the host-virus association can in some settings be slanted towards ART-free viral remission.
5. Deeks, S. G. et al. International AIDS Society global scientific strategy: towards an HIV cure 2016. *Nat. Med.* **22**, 839–850 (2016).
6. Venter, W. D. F. et al. Dolutegravir plus two different prodrugs of tenofovir to treat HIV. *N. Engl. J. Med.* **381**, 803–815 (2019).
7. The NAMSA ANRS 12313 Study Group. Dolutegravir-based or low-dose efavirenz-based regimen for the treatment of HIV-1. *N. Engl. J. Med.* **381**, 816–826 (2019).
8. Rueda, S. et al. Examining the associations between HIV-related stigma and health outcomes in people living with HIV/AIDS: a series of meta-analyses. *BMJ Open* **6**, e011453 (2016).
9. Katz, I. T. et al. Impact of HIV-related stigma on treatment adherence: systematic review and meta-synthesis. *J. Int. AIDS Soc.* **16**, 18640 (2013).
10. Chu, C. E. et al. Exploring the social meaning of curing HIV: a qualitative study of people who inject drugs in Guangzhou, China. *AIDS Res. Hum. Retroviruses* **31**, 78–84 (2015).
11. Eisinger, R. W., Dieffenbach, C. W. & Fauci, A. S. HIV viral load and transmissibility of HIV infection: undetectable equals untransmittable. *J. Am. Med. Assoc.* **321**, 451–452 (2019).
12. Rodger, A. J. et al. Risk of HIV transmission through condomless sex in serodifferent gay couples with the HIV-positive partner taking suppressive antiretroviral therapy (PARTNER): final results of a multicentre, prospective, observational study. *Lancet* **393**, 2428–2438 (2019).

13. UNAIDS. *Global HIV & AIDS Statistics — 2019 Fact Sheet* <https://www.unaids.org/en/resources/fact-sheet> (2019)
14. Beacroft, L. & Hallett, T. B. The potential impact of a “curative intervention” for HIV: a modelling study. *Glob. Health Res. Policy* **4**, 18 (2019).
15. Cuadros, D. F. et al. Towards UNAIDS Fast-Track goals: targeting priority geographic areas for HIV prevention and care in Zimbabwe. *AIDS* **33**, 305–314 (2019).
16. GBD 2017 HIV collaborators. Global, regional, and national incidence, prevalence, and mortality of HIV, 1980–2017, and forecasts to 2030, for 195 countries and territories: a systematic analysis for the Global Burden of Diseases, Injuries, and Risk Factors Study 2017. *Lancet HIV* **6**, e831–e859 (2019).
17. Moyo, S. et al. Cross-sectional estimates revealed high HIV incidence in Botswana rural communities in the era of successful ART scale-up in 2013–2015. *PLoS ONE* **13**, e0204840 (2018).
18. UNAIDS. *Fast-Track Update On Investments Needed In The AIDS Response* https://www.unaids.org/sites/default/files/media_asset/UNAIDS_Reference_FastTrack_Update_on_investments_en.pdf (2016).
19. Gelpi, A. & Tucker, J. D. A cure at last? Penicillin’s unintended consequences on syphilis control, 1944–1964. *Sex. Transm. Infect.* **91**, 70 (2015).
20. Finzi, D. et al. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**, 1295–1300 (1997).
21. Wong, J. K. et al. Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science* **278**, 1291–1295 (1997).
22. Chun, T. W. et al. Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proc. Natl Acad. Sci. USA* **94**, 13193–13197 (1997).
23. Wiegand, A. et al. Single-cell analysis of HIV-1 transcriptional activity reveals expression of proviruses in expanded clones during ART. *Proc. Natl Acad. Sci. USA* **114**, E3659–E3668 (2017).
24. Cohn, L. B. et al. Clonal CD4⁺ T cells in the HIV-1 latent reservoir display a distinct gene profile upon reactivation. *Nat. Med.* **24**, 604–609 (2018).
- Using single-cell analyses, the transcriptional profile of HIV-infected CD4⁺ T cells was characterized, revealing the expression of pathways that suppress HIV transcription.**
25. Kuo, H.-H. et al. Anti-apoptotic protein BIRC5 maintains survival of HIV-1-infected CD4⁺ T cells. *Immunity* **48**, 1183–1194 (2018).
26. Yukl, S. A. et al. HIV latency in isolated patient CD4⁺ T cells may be due to blocks in HIV transcriptional elongation, completion, and splicing. *Sci. Transl. Med.* **10**, eaap9927 (2018).
- Multiple blocks to HIV transcription exist in infected CD4⁺ T cells; it may be necessary to overcome each using diverse approaches before sufficient amounts of HIV protein are made for the cell to be recognized and eliminated.**
27. Ho, Y. C. et al. Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell* **155**, 540–551 (2013).
- Most (>90%) integrated HIV genomes are defective and unable to support HIV replication, complicating the measurement of the inducible replication-competent viral reservoir.**
28. Hosmane, N. N. et al. Proliferation of latently infected CD4⁺ T cells carrying replication-competent HIV-1: potential role in latent reservoir dynamics. *J. Exp. Med.* **214**, 959–972 (2017).
29. Descours, B. et al. CD32a is a marker of a CD4 T-cell HIV reservoir harbouring replication-competent proviruses. *Nature* **543**, 564–567 (2017).
30. Chomont, N. et al. HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. *Nat. Med.* **15**, 893–900 (2009).
31. Fromentin, R. et al. CD4⁺ T cells expressing PD-1, TIGIT and LAG-3 contribute to HIV persistence during ART. *PLoS Pathog.* **12**, e1005761 (2016).
32. Banga, R. et al. PD-1⁺ and follicular helper T cells are responsible for persistent HIV-1 transcription in treated aviremic individuals. *Nat. Med.* **22**, 754–761 (2016).
- During effective antiretroviral therapy, transcriptionally active HIV is enriched in PD-1-expressing T follicular helper cells that reside in lymph nodes.**
33. McGary, C. S. et al. CTLA-4⁺PD-1⁺ memory CD4⁺ T cells critically contribute to viral persistence in antiretroviral therapy-suppressed, SIV-infected rhesus macaques. *Immunity* **47**, 776–788 (2017).
34. Chew, G. M. et al. TIGIT marks exhausted T cells, correlates with disease progression, and serves as a target for immune restoration in HIV and SIV infection. *PLoS Pathog.* **12**, e1005349 (2016).
35. Hiener, B. et al. Identification of genetically intact HIV-1 proviruses in specific CD4⁺ T cells from effectively treated participants. *Cell Rep.* **21**, 813–822 (2017).
36. Hogan, L. E. et al. Increased HIV-1 transcriptional activity and infectious burden in peripheral blood and gut-associated CD4⁺ T cells expressing CD30. *PLoS Pathog.* **14**, e1006856 (2018).
37. Pardons, M. et al. Single-cell characterization and quantification of translation-competent viral reservoirs in treated and untreated HIV infection. *PLoS Pathog.* **15**, e1007619 (2019).
38. Khoury, G. et al. Persistence of integrated HIV DNA in CXCR3⁺CCR6⁺ memory CD4⁺ T cells in HIV-infected individuals on antiretroviral therapy. *AIDS* **30**, 1511–1520 (2016).
39. Lee, G. Q. et al. Clonal expansion of genome-intact HIV-1 in functionally polarized Th1 CD4⁺ T cells. *J. Clin. Invest.* **127**, 2689–2696 (2017).
40. Wacleche, V. S. et al. New insights into the heterogeneity of Th17 subsets contributing to HIV-1 persistence during antiretroviral therapy. *Retrovirology* **13**, 59 (2016).
41. Wang, Z. et al. Expanded cellular clones carrying replication-competent HIV-1 persist, wax, and wane. *Proc. Natl Acad. Sci. USA* **115**, E2575–E2584 (2018).
42. Wagner, T. A. et al. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science* **345**, 570–573 (2014).
43. Maldarelli, F. et al. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345**, 179–183 (2014).
- As demonstrated in these two studies^{42,43}, proliferation of memory T cells is the main mechanism by which the reservoir is maintained indefinitely; during long-term therapy, the population of infected cells becomes increasingly clonal with integration of defective HIV genomes in genes associated with cell growth (including oncogenes) and/or within intergenic regions or silent genes.**
44. Cohn, L. B. et al. HIV-1 integration landscape during latent and active infection. *Cell* **160**, 420–432 (2015).
45. Einkauf, K. B. et al. Intact HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy. *J. Clin. Invest.* **129**, 988–998 (2019).
46. Battivelli, E. et al. Distinct chromatin functional states correlate with HIV latency reactivation in infected primary CD4⁺ T cells. *eLife* **7**, e34655 (2018).
47. Pinzone, M. R. et al. Longitudinal HIV sequencing reveals reservoir expression leading to decay which is obscured by clonal expansion. *Nat. Commun.* **10**, 728 (2019).
48. Lee, G. Q. et al. HIV-1 DNA sequence diversity and evolution during acute subtype C infection. *Nat. Commun.* **10**, 2737 (2019).
49. Huang, S. H. et al. Latent HIV reservoirs exhibit inherent resistance to elimination by CD8⁺ T cells. *J. Clin. Invest.* **128**, 876–889 (2018).
50. Fletcher, C. V. et al. Persistent HIV-1 replication is associated with lower antiretroviral drug concentrations in lymphatic tissues. *Proc. Natl Acad. Sci. USA* **111**, 2307–2312 (2014).
51. Kearney, M. F. et al. Lack of detectable HIV-1 molecular evolution during suppressive antiretroviral therapy. *PLoS Pathog.* **10**, e1004010 (2014).
52. Shan, L. et al. Stimulation of HIV-1-specific cytolytic T lymphocytes facilitates elimination of latent viral reservoir after virus reactivation. *Immunity* **36**, 491–501 (2012).
53. Deng, K. et al. Broad CTL response is required to clear latent HIV-1 due to dominance of escape mutations. *Nature* **517**, 381–385 (2015).
54. Migueles, S. A. et al. Defective human immunodeficiency virus-specific CD8⁺ T-cell polyfunctionality, proliferation, and cytotoxicity are not restored by antiretroviral therapy. *J. Virol.* **83**, 11876–11889 (2009).
55. Peretz, Y. et al. CD160 and PD-1 co-expression on HIV-specific CD8 T cells defines a subset with advanced dysfunction. *PLoS Pathog.* **8**, e1002840 (2012).
56. Hersperger, A. R. et al. Increased HIV-specific CD8⁺ T-cell cytotoxic potential in HIV elite controllers is associated with T-bet expression. *Blood* **117**, 3799–3808 (2011).
57. Tauriainen, J. et al. Perturbed CD8⁺ T cell TIGIT/CD226/PVR axis despite early initiation of antiretroviral treatment in HIV infected individuals. *Sci. Rep.* **7**, 40354 (2017).
58. Takata, H. et al. Delayed differentiation of potent effector CD8⁺ T cells reducing viremia and reservoir seeding in acute HIV infection. *Sci. Transl. Med.* **9**, eaag1809 (2017).
59. Ndhlovu, Z. M. et al. Augmentation of HIV-specific T cell function by immediate treatment of hyperacute HIV-1 infection. *Sci. Transl. Med.* **11**, eaau0528 (2019).
60. Ndhlovu, Z. M. et al. Magnitude and kinetics of CD8⁺ T cell activation during hyperacute HIV infection impact viral set point. *Immunity* **43**, 591–604 (2015).
- Studies^{58–60} have shown that acute HIV infection is associated with a large CD8⁺ T cell response, which is initially effective but rapidly becomes less functional.**
61. Yukl, S. A. et al. Challenges in detecting HIV persistence during potentially curative interventions: a study of the Berlin patient. *PLoS Pathog.* **9**, e1003347 (2013).
62. Colby, D. J. et al. Rapid HIV RNA rebound after antiretroviral treatment interruption in persons durably suppressed in Fiebig I acute HIV infection. *Nat. Med.* **24**, 923–926 (2018).
63. Henrich, T. J. et al. HIV-1 persistence following extremely early initiation of antiretroviral therapy (ART) during acute HIV-1 infection: an observational study. *PLoS Med.* **14**, e1002417 (2017).
- Studies^{62,63} have demonstrated that even with highly effective interventions that result in multi-log reductions in the reservoir, a small undetectable reservoir of replication-competent HIV can persist and rebound many months after ART is interrupted.**
64. Henrich, T. J. et al. Antiretroviral-free HIV-1 remission and viral rebound after allogeneic stem cell transplantation: report of 2 cases. *Ann. Intern. Med.* **161**, 319–327 (2014).
65. Persaud, D. et al. Absence of detectable HIV-1 viremia after treatment cessation in an infant. *N. Engl. J. Med.* **369**, 1828–1835 (2013).
66. Fidler, S. et al. A pilot evaluation of whole blood finger-prick sampling for point-of-care HIV viral load measurement: the UNICORN study. *Sci. Rep.* **7**, 13658 (2017).
67. Bruner, K. M. et al. A quantitative approach for measuring the reservoir of latent HIV-1 proviruses. *Nature* **566**, 120–125 (2019).
68. Estes, J. D. et al. Defining total-body AIDS-virus burden with implications for curative strategies. *Nat. Med.* **23**, 1271–1276 (2017).
69. Santangelo, P. J. et al. Whole-body immunoPET reveals active SIV dynamics in viremic and antiretroviral therapy-treated macaques. *Nat. Methods* **12**, 427–432 (2015).
70. Keating, S. M. et al. HIV antibody level as a marker of HIV persistence and low-level viral replication. *J. Infect. Dis.* **216**, 72–81 (2017).
71. Rothenberger, M. K. et al. Large number of rebounding/founder HIV variants emerge from multifocal infection in lymphatic tissues after treatment interruption. *Proc. Natl Acad. Sci. USA* **112**, E1126–E1134 (2015).
72. Sneller, M. C. et al. A randomized controlled safety/efficacy trial of therapeutic vaccination in HIV-infected individuals who initiated antiretroviral therapy early in infection. *Sci. Transl. Med.* **9**, eaan8848 (2017).
73. Lelièvre, J.-D. & Hocqueloux, L. Unintended HIV-1 transmission to a sex partner in a study of a therapeutic vaccine candidate. *J. Infect. Dis.* **220**, S5–S6 (2019).
74. Julg, B. et al. Recommendations for analytical antiretroviral treatment interruptions in HIV research trials—report of a consensus meeting. *Lancet HIV* **6**, e259–e268 (2019).
75. Phillips, A. N. et al. Identifying key drivers of the impact of an HIV cure intervention in sub-Saharan Africa. *J. Infect. Dis.* **214**, 73–79 (2016).
76. Paltiel, A. D. et al. Setting performance standards for a cost-effective human immunodeficiency virus cure strategy in South Africa. *Open Forum Infect. Dis.* **4**, ofx081 (2017).
77. Dong, K. L. et al. Detection and treatment of Fiebig stage I HIV-1 infection in young at-risk women in South Africa: a prospective cohort study. *Lancet HIV* **5**, e35–e44 (2018).
78. Okoye, A. A. et al. Early antiretroviral therapy limits SIV reservoir establishment to delay or prevent post-treatment viral rebound. *Nat. Med.* **24**, 1430–1440 (2018).
79. Violari, A. et al. A child with perinatal HIV infection and long-term sustained virological control following antiretroviral treatment cessation. *Nat. Commun.* **10**, 412 (2019).
80. Frange, P. et al. HIV-1 virological remission lasting more than 12 years after interruption of early antiretroviral therapy in a perinatally infected teenager enrolled in the French ANRS EPF-CO10 paediatric cohort: a case report. *Lancet HIV* **3**, e49–e54 (2016).
81. Holt, N. et al. Human hematopoietic stem/progenitor cells modified by zinc-finger nucleases targeted to CCR5 control HIV-1 in vivo. *Nat. Biotechnol.* **28**, 839–847 (2010).

82. Peterson, C. W. et al. Differential impact of transplantation on peripheral and tissue-associated viral reservoirs: implications for HIV gene therapy. *PLoS Pathog.* **14**, e1006956 (2018).
 83. Haworth, K. G., Peterson, C. W. & Kiem, H. P. CCR5-edited gene therapies for HIV cure: closing the door to viral entry. *Cytherapy* **19**, 1325–1338 (2017).
 84. Yin, H., Kauffman, K. J. & Anderson, D. G. Delivery technologies for genome editing. *Nat. Rev. Drug Discov.* **16**, 387–399 (2017).
 85. Davenport, M. P. et al. Functional cure of HIV: the scale of the challenge. *Nat. Rev. Immunol.* **19**, 45–54 (2019).
 86. Kordelas, L. et al. Shift of HIV tropism in stem-cell transplantation with CCR5 Delta32 mutation. *N. Engl. J. Med.* **371**, 880–882 (2014).
 87. Dash, P. K. et al. Sequential LASER ART and CRISPR treatments eliminate HIV-1 in a subset of infected humanized mice. *Nat. Commun.* **10**, 2753 (2019).
 88. Wang, G., Zhao, N., Berkhout, B. & Das, A. T. A combinatorial CRISPR–Cas9 attack on HIV-1 DNA extinguishes all infectious provirus in infected T cell cultures. *Cell Rep.* **17**, 2819–2826 (2016).
 89. Martinez-Navio, J. M. et al. Adeno-associated virus delivery of anti-HIV monoclonal antibodies can drive long-term virologic suppression. *Immunity* **50**, 567–575 (2019).
 90. Priddy, F. H. et al. Adeno-associated virus vectored immunoprophylaxis to prevent HIV in healthy adults: a phase 1 randomised controlled trial. *Lancet HIV* **6**, e230–e239 (2019).
 91. Deeks, S. G. et al. A phase II randomized study of HIV-specific T-cell gene therapy in subjects with undetectable plasma viremia on combination antiretroviral therapy. *Mol. Ther.* **5**, 788–797 (2002).
 92. June, C. H., O'Connor, R. S., Kawalekar, O. U., Ghassemi, S. & Milone, M. C. CAR T cell immunotherapy for human cancer. *Science* **359**, 1361–1365 (2018).
 93. Zhen, A. et al. Long-term persistence and function of hematopoietic stem cell-derived chimeric antigen receptor T cells in a nonhuman primate model of HIV/AIDS. *PLoS Pathog.* **13**, e1006753 (2017).
 94. Leibman, R. S. et al. Supraphysiologic control over HIV-1 replication mediated by CD8 T cells expressing a re-engineered CD4-based chimeric antigen receptor. *PLoS Pathog.* **13**, e1006613 (2017).
 95. Anthony-Gonda, K. et al. Multispecific anti-HIV duoCAR-T cells display broad in vitro antiviral activity and potent in vivo elimination of HIV-infected cells in a humanized mouse model. *Sci. Transl. Med.* **11**, eaav5685 (2019).
 96. Archin, N. M. et al. Administration of vorinostat disrupts HIV-1 latency in patients on antiretroviral therapy. *Nature* **487**, 482–485 (2012).
 97. Elliott, J. H. et al. Activation of HIV transcription with short-course vorinostat in HIV-infected patients on suppressive antiretroviral therapy. *PLoS Pathog.* **10**, e1004473 (2014).
 98. Elliott, J. H. et al. Short-term administration of disulfiram for reversal of latent HIV infection: a phase 2 dose-escalation study. *Lancet HIV* **2**, e520–e529 (2015).
 99. Rasmussen, T. A. et al. Panobinostat, a histone deacetylase inhibitor, for latent-virus reactivation in HIV-infected patients on suppressive antiretroviral therapy: a phase 1/2, single group, clinical trial. *Lancet HIV* **1**, e13–e21 (2014).
 100. Søgaard, O. S. et al. The depsi-peptide romidepsin reverses HIV-1 latency in vivo. *PLoS Pathog.* **11**, e1005142 (2015).
 101. Lim, S.-Y. et al. TLR7 agonists induce transient viremia and reduce the viral reservoir in SIV-infected rhesus macaques on antiretroviral therapy. *Sci. Transl. Med.* **10**, eaao4521 (2018).
 102. Vibholm, L. et al. Short-course Toll-like receptor 9 agonist treatment impacts innate immunity and plasma viremia in individuals with human immunodeficiency virus infection. *Clin. Infect. Dis.* **64**, 1686–1695 (2017).
 103. Laird, G. M. et al. Ex vivo analysis identifies effective HIV-1 latency-reversing drug combinations. *J. Clin. Invest.* **125**, 1901–1912 (2015).
 104. Macedo, A. B. et al. Dual TLR2 and TLR7 agonists as HIV latency-reversing agents. *JCI Insight* **3**, e122673 (2018).
 105. Rochat, M. A., Schlaepfer, E. & Speck, R. F. Promising role of Toll-like receptor 8 agonist in concert with prostratin for activation of silent HIV. *J. Virol.* **91**, e02084-16 (2017).
 106. Hill, A. L., Rosenbloom, D. I., Fu, F., Nowak, M. A. & Siliciano, R. F. Predicting the outcomes of treatment to eradicate the latent reservoir for HIV-1. *Proc. Natl Acad. Sci. USA* **111**, 13475–13480 (2014).
 107. Mousseau, G. et al. The Tat inhibitor didehydro-cortistatin A prevents HIV-1 reactivation from latency. *mBio* **6**, e00465-15 (2015).
 108. Kessing, C. F. et al. In vivo suppression of HIV rebound by didehydro-cortistatin A, a “block-and-lock” strategy for HIV-1 treatment. *Cell Rep.* **21**, 600–611 (2017).
 109. Barr, E. L. & Jefferys, R. A landscape analysis of HIV cure-related clinical trials and observational studies in 2018. *J. Virus Eradication* http://viruseradication.com/journal-details/A_landscape_analysis_of_HIV_cure-related_clinical_trials_and_observational_studies_in_2018/ (2019).
 110. Conway, J. M. & Perelson, A. S. Post-treatment control of HIV infection. *Proc. Natl Acad. Sci. USA* **112**, 5467–5472 (2015).
 111. Casazza, J. P. et al. Therapeutic vaccination expands and improves the function of the HIV-specific memory T-cell repertoire. *J. Infect. Dis.* **207**, 1829–1840 (2013).
 112. Hansen, S. G. et al. Broadly targeted CD8⁺ T cell responses restricted by major histocompatibility complex E. *Science* **351**, 714–720 (2016).
 113. Hansen, S. G. et al. Immune clearance of highly pathogenic SIV infection. *Nature* **502**, 100–104 (2013).
 114. Hessel, A. J. et al. Early short-term treatment with neutralizing human monoclonal antibodies halts SHIV infection in infant macaques. *Nat. Med.* **22**, 362–368 (2016).
 115. Borducchi, E. N. et al. Ad26/MVA therapeutic vaccination with TLR7 stimulation in SIV-infected rhesus monkeys. *Nature* **540**, 284–287 (2016).
 116. Borducchi, E. N. et al. Antibody and TLR7 agonist delay viral rebound in SHIV-infected monkeys. *Nature* **563**, 360–364 (2018).
- In these two related non-human primate studies^{116,118}, a combination of immune therapies resulted in enhanced immune control and, in some cases, possible elimination of SIV/SHIV; these strategies are now being tested in people.**
117. Nishimura, Y. et al. Early antibody therapy can induce long-lasting immunity to SHIV. *Nature* **543**, 559–563 (2017).
 118. Mendoza, P. et al. Combination therapy with anti-HIV-1 antibodies maintains viral suppression. *Nature* **561**, 479–484 (2018).
- These two related studies^{117,118} demonstrate that broadly neutralizing antibodies administered during a period of acute SHIV infection of non-human primates or immediately after interruption of ART in HIV-infected humans potentially induce long-term immune-mediated control (the ‘vaccinal effect’); in the case of SHIV-infected non-human primates, such suppression was shown to be mediated by CD8⁺ T cells.**
119. Xu, L. et al. Trispecific broadly neutralizing HIV antibodies mediate potent SHIV protection in macaques. *Science* **358**, 85–90 (2017).
 120. Sung, J. A. et al. Dual-affinity re-targeting proteins direct T cell-mediated cytotoxicity of latently HIV-infected cells. *J. Clin. Invest.* **125**, 4077–4090 (2015).
 121. Leal, L. et al. New challenges in therapeutic vaccines against HIV infection. *Expert Rev. Vaccines* **16**, 587–600 (2017).
 122. Gaiha, G. D. et al. Structural topology defines protective CD8⁺ T cell epitopes in the HIV proteome. *Science* **364**, 480–484 (2019).
 123. Fromentin, R. et al. PD-1 blockade potentiates HIV latency reversal ex vivo in CD4⁺ T cells from ART-suppressed individuals. *Nat. Commun.* **10**, 814 (2019).
 124. Gay, C. L. et al. Clinical trial of the anti-PD-L1 antibody BMS-936559 in HIV-1 infected participants on suppressive antiretroviral therapy. *J. Infect. Dis.* **215**, 1725–1733 (2017).
 125. Webb, G. M. et al. The human IL-15 superagonist ALT-803 directs SIV-specific CD8⁺ T cells into B-cell follicles. *Blood Adv.* **2**, 76–84 (2018).
 126. Petrovas, C. et al. Follicular CD8 T cells accumulate in HIV infection and can kill infected cells in vitro via bispecific antibodies. *Sci. Transl. Med.* **9**, eaag2285 (2017).
 127. Ananworanich, J. & Barré-Sinoussi, F. Is it time to abandon single intervention cure trials? *Lancet HIV* **2**, e410–e411 (2015).
 128. Leth, S. et al. Combined effect of Vacc-4x, recombinant human granulocyte macrophage colony-stimulating factor vaccination, and romidepsin on the HIV-1 reservoir (REDUC): a single-arm, phase 1B/2A trial. *Lancet HIV* **3**, e463–e472 (2016).
 129. Rossouw, T., Tucker, J. D., van Zyl, G. U., Sikwesi, K. & Godfrey, C. Barriers to HIV remission research in low- and middle-income countries. *J. Int. AIDS Soc.* **20**, 21521 (2017).
 130. Kityo, C. et al. Lymphoid tissue fibrosis is associated with impaired vaccine responses. *J. Clin. Invest.* **128**, 2763–2773 (2018).
 131. Scully, E. P. et al. Sex-based differences in human immunodeficiency virus type 1 reservoir activity and residual immune activation. *J. Infect. Dis.* **219**, 1084–1094 (2019).
- Acknowledgements** T.N. is supported in part by grants from the Bill & Melinda Gates Foundation, Gilead Sciences (grant no. 00406), the International AIDS Vaccine Initiative (IAVI) (UKZNRSA1001), the NIAID (R37AI067073) and the South African Department of Science and Technology through the National Research Foundation. T.N. is also partially supported through the sub-Saharan African Network for TB/HIV Research Excellence (SANTHE), a DELTAS Africa Initiative (grant no. DEL-15-006). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS) Alliance for Accelerating Excellence in Science in Africa (AESA) and is supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (grant no. 107752/Z/15/Z) and the UK government. J.M.M. is an employee of the Bill & Melinda Gates Foundation. S.G.D. is supported by the amfAR Institute for HIV Cure Research (amfAR 109301), the Delaney AIDS Research Enterprise (DARE; A127966), and the NIAID (K24 AI069994). We thank W. Greene for assistance with Fig. 1. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust, the UK government, the Bill & Melinda Gates Foundation or amfAR.
- Author contributions** T.N. and S.G.D. produced the first draft. J.M.M. provided substantial additions and edits. All three edited and approved the final version.
- Competing interests** S.G.D. receives grant support from Gilead, Merck and ViiV. He is a member of the scientific advisory boards for BryoLogix and Enochian Biosciences and has consulted for AbbVie, Biotron and Eli Lilly. T.N. receives grant support from Gilead.
- Additional information**
Correspondence and requests for materials should be addressed to S.G.D.
Reprints and permissions information is available at <http://www.nature.com/reprints>.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© Springer Nature Limited 2019

A statistical solution to the chaotic, non-hierarchical three-body problem

<https://doi.org/10.1038/s41586-019-1833-8>

Nicholas C. Stone^{1,2,3*} & Nathan W. C. Leigh^{4,5}

Received: 4 February 2019

Accepted: 28 September 2019

Published online: 18 December 2019

The three-body problem is arguably the oldest open question in astrophysics and has resisted a general analytic solution for centuries. Various implementations of perturbation theory provide solutions in portions of parameter space, but only where hierarchies of masses or separations exist. Numerical integrations¹ show that bound, non-hierarchical triple systems of Newtonian point particles will almost² always disintegrate into a single escaping star and a stable bound binary^{3,4}, but the chaotic nature of the three-body problem⁵ prevents the derivation of tractable⁶ analytic formulae that deterministically map initial conditions to final outcomes. Chaos, however, also motivates the assumption of ergodicity^{7–9}, implying that the distribution of outcomes is uniform across the accessible phase volume. Here we report a statistical solution to the non-hierarchical three-body problem that is derived using the ergodic hypothesis and that provides closed-form distributions of outcomes (for example, binary orbital elements) when given the conserved integrals of motion. We compare our outcome distributions to large ensembles of numerical three-body integrations and find good agreement, so long as we restrict ourselves to ‘resonant’ encounters¹⁰ (the roughly 50 per cent of scatterings that undergo chaotic evolution). In analysing our scattering experiments, we identify ‘scrambles’ (periods of time in which no pairwise binaries exist) as the key dynamical state that ergodizes a non-hierarchical triple system. The generally super-thermal distributions of survivor binary eccentricity that we predict have notable applications to many astrophysical scenarios. For example, non-hierarchical triple systems produced dynamically in globular clusters are a primary formation channel for black-hole mergers^{11–13}, but the rates and properties^{14,15} of the resulting gravitational waves depend on the distribution of post-disintegration eccentricities.

The three-body problem is a prototypical example of deterministic chaos⁵, in that tiny perturbations in the initial conditions (or errors in numerical integration) lead to exponentially divergent outcomes¹⁶. Chaotic systems often ‘forget’ their initial conditions (aside from integrals of motion), although this is by no means guaranteed—indeed, the topology of the chaotic three-body problem does contain islands of regularity^{17,18}. Nonetheless, to a first approximation, it is reasonable to estimate the probability of different outcomes by invoking the ergodic hypothesis^{7,19} and to assume that non-hierarchical triples will uniformly explore the phase-space volume accessible to them⁸. In this way, we may turn the chaotic nature of the three-body problem^{5,16}—which has so far frustrated general, deterministic, analytic mappings from one set of initial conditions to one set of outcomes—into a tool that simplifies the mapping from distributions of initial conditions to distributions of outcomes.

We consider the generic outcome of the non-hierarchical Newtonian three-body problem: a single escaper star with mass m_s departs from a surviving binary with mass $m_b = m_a + m_b$, where m_a and m_b are the

component masses. In Fig. 1 we illustrate this scenario, using both a direct numerical integration of the equations of motion and a schematic diagram of a metastable triple at the moment of breakup. At the time of disintegration, the binary components are separated by a distance \mathbf{r} and have relative momentum \mathbf{p} , and the escaper is separated from the binary centre of mass by \mathbf{r}_s and is moving with relative momentum \mathbf{p}_s . The total energy and angular momentum of the system, inherited from the initial conditions and preserved through a period of chaotic three-body interactions, are E_0 and \mathbf{L}_0 , respectively. For convenience, we define the additional masses $M = m_s + m_b$, $m = m_b m_s / M$ and $\mathcal{M} = m_a m_b / m_b$. The total accessible phase volume for this system is that of an eight-dimensional hypersurface⁸:

$$\sigma = \int \dots \int \delta(E_B + E_s - E_0) \delta(\mathbf{L}_B + \mathbf{L}_s - \mathbf{L}_0) d\mathbf{r} d\mathbf{p} d\mathbf{r}_s d\mathbf{p}_s \quad (1)$$

shaped by the requirements of energy and angular-momentum conservation for both the elliptic orbit of the surviving binary (E_B , \mathbf{L}_B) and the hyperbolic orbit between the binary and the escaper (E_s , \mathbf{L}_s). In

¹Columbia Astrophysics Laboratory, Columbia University, New York, NY, USA. ²Racah Institute of Physics, The Hebrew University, Jerusalem, Israel. ³Department of Astronomy, University of Maryland, College Park, MD, USA. ⁴Departamento de Astronomía, Facultad de Ciencias Físicas y Matemáticas, Universidad de Concepción, Concepción, Chile. ⁵Department of Astrophysics, American Museum of Natural History, New York, NY, USA. *e-mail: nicholas.stone@mail.huji.ac.il

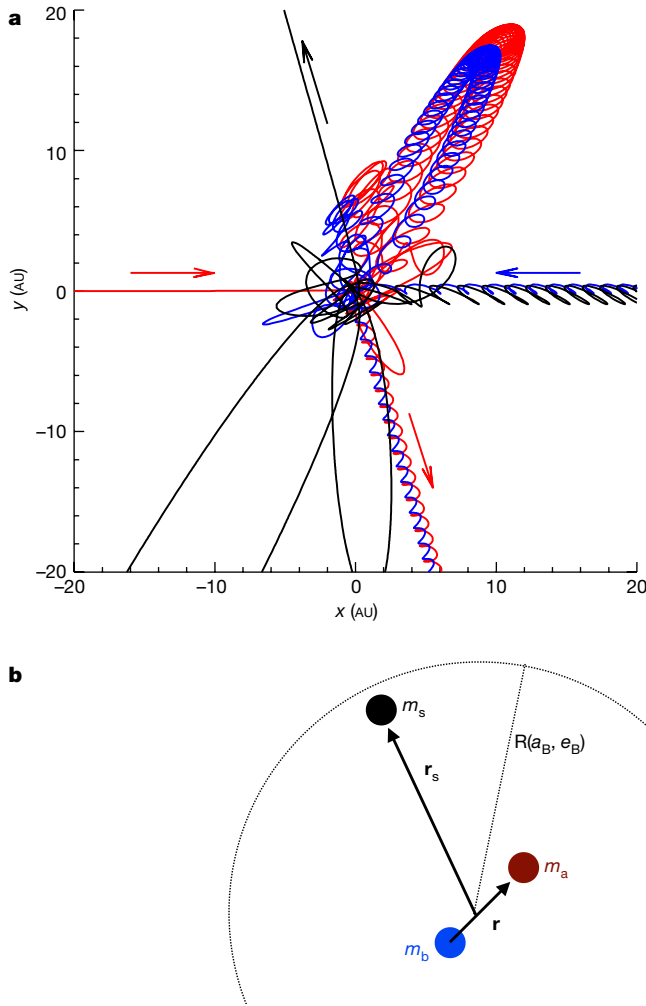


Fig. 1 | Non-hierarchical three-body scatterings. **a**, Two-dimensional projection of an equal-mass resonant scattering encounter, where an interloper star (red) encounters a binary (blue and black). The resonant interaction unfolds over several dynamical times before the system disintegrates in a partner swap. **b**, Schematic illustration of the metastable triple at the moment of disintegration.

equation (1), δ represents the Dirac delta function. Given a microcanonical ensemble of non-hierarchical triples with different initial conditions but identical integrals of motion and mass combinations, the outcome states (after breakup) will—assuming ergodicity—uniformly populate the phase volume that is accessible at the moment of disintegration. This ensemble is microcanonical in the sense that each three-body system is isolated from external sources of heat, but is unusual in its low particle number⁷.

We evaluate this integral at the moment of disintegration, which we idealize as occurring anywhere inside a ‘strong interaction region’ of radius $R(E_B, L_B, C_B)$, where $C_B = \hat{\mathbf{L}}_B \cdot \hat{\mathbf{L}}_0$. Canonical transformations to elliptic/hyperbolic Delaunay elements facilitate the integration (see Supplementary Information) and yield a phase volume of

$$\sigma = \frac{2\pi^4 G^2 M^{5/2} m_B}{(m_a m_b m_s)^{3/2}} \iiint \frac{L_B dE_B dL_B dC_B}{L_s (-E_B)^{3/2} (E_0 - E_B)^{3/2}} \times \left[\sqrt{\frac{2M(E_0 - E_B)}{G^2 m_s^3 m_B^3}} \sqrt{2m(E_0 - E_B)R^2 + 2GMm^2R - L_s^2} - \operatorname{acosh} \left(\frac{1 + [2(E_0 - E_B)R/(Gm_s m_B)]}{\sqrt{1 + [2M(E_0 - E_B)L_s^2/(G^2 m_s^3 m_B^3)]}} \right) \right] \quad (2)$$

where G is Newton’s gravitational constant. For brevity, we have re-inserted the angular momentum of the escaping star, $L_s^2(L_B, C_B) \equiv L_B^2(1 - C_B^2) + (L_B C_B - L_0)^2$. σ is a phase volume and the integrand of equation (2) is a trivariate outcome distribution representing the differential probability of finding a disintegrating metastable triple in a volume $dE_B dL_B dC_B$: the microcanonical ensemble for survivor binaries produced in the non-hierarchical three-body problem (other—angular—binary orbital elements are distributed uniformly). Therefore, specification of the total energy E_0 and the total angular momentum \mathbf{L}_0 suffices to describe the distribution of outcomes in non-hierarchical triple systems, even if this information alone cannot deterministically specify how one individual outcome follows from one set of initial conditions. Conservation of E_0 and \mathbf{L}_0 means that the trivariate outcome distribution in equation (2) can be mapped one to one to the distribution of escaper properties. Equation (2) makes fewer simplifying assumptions than did past ergodic analyses of the general three-body problem^{8,9,20,21}, and its outcome distributions are qualitatively different.

We marginalize over L_B and C_B to compute the distribution of outcome energies, $d\sigma/dE_B$. In this and all remaining calculations, we assume that the strong interaction region is a dimensionless multiple of the time-averaged binary size, that is, $R = \alpha a_B(1 + e_B)$, where $\alpha \approx 1$ is a dimensionless constant (see Extended Data Figs. 1–3 and Supplementary Information for more details). In an $L_0 = 0$ ensemble, this is $d\sigma/dE_B \propto |E_B|^{-7/2}$, extending to $|E_B| \rightarrow \infty$. Conversely, when L_0 is large, the ergodic energy distribution is slightly steeper, changing roughly as $d\sigma/dE_B \propto |E_B|^{-4}$, but only up to a maximum energy of $|E_{\max}| \propto L_0^2$; larger outcome energies are prohibited by angular-momentum conservation. The energy distribution that we calculate differs from past estimates determined assuming detailed balance¹⁰, demonstrating that a population of binaries engaging in resonant three-body interactions with a thermal bath of single stars cannot achieve detailed balance, so long as their outcomes are ergodically distributed.

We likewise integrate to find the marginal outcome distributions in angular momentum (which we represent in terms of binary eccentricity e_B , as $d\sigma/de_B$) and inclination ($d\sigma/dC_B$). In contrast to the usual (although not universal²²) expectation of a thermal eccentricity distribution, $d\sigma/de_B = 2e_B$, we find a mildly super-thermal eccentricity distribution for large L_0 : $d\sigma/de_B = \frac{6}{5}e_B(1 + e_B)$. This radial orbit bias is a geometric effect arising from the larger average interaction cross-section of a highly eccentric binary, the apocentre of which is twice as large as that of a circular binary of equal energy. In the low- L_0 limit, the ergodic distribution of survivor eccentricities is highly super-thermal, with $d\sigma/de_B \propto e_B(1 + e_B)/\sqrt{1 - e_B^2}$ when $L_0 = 0$. There is a strong bias towards producing nearly radial binaries as a consequence of angular-momentum starvation: whereas a low- L_0 ensemble of non-hierarchical triples may produce a quasi-circular survivor binary, doing so requires substantial fine-tuning of the angle and velocity of the escaper, and is therefore disfavoured. Similar phase volume considerations explain the strong bias towards prograde ($0 < C_B \leq 1$) orbits predicted by equation (2) when marginalized into $d\sigma/dC_B$. More detailed explorations of the ergodic $d\sigma/dE_B$, $d\sigma/de_B$ and $d\sigma/dC_B$ distributions are shown in Extended Data Figs. 1, 2, 3, respectively, as well as in Supplementary Information.

Our outcome distribution, $d\sigma/(dE_B dL_B dC_B)$, was derived with several assumptions, most notably: (i) the ergodic hypothesis, (ii) instantaneous disintegration and (iii) a specific parameterization of the ‘strong interaction region’ defining the limits of integration. It should therefore be tested against ensembles of numerical scattering experiments. We have explored the ergodicity of non-hierarchical triples in the equal-mass limit by using the FEWBODY numerical scattering code to run three ensembles of different binary–single scattering experiments (see Extended Data Table 1). Each ensemble has roughly $N \approx 10^5$ runs with constant E_0 and L_0 , but otherwise random initial conditions (we initialize our binary–single scatterings with zero impact parameter, so we can

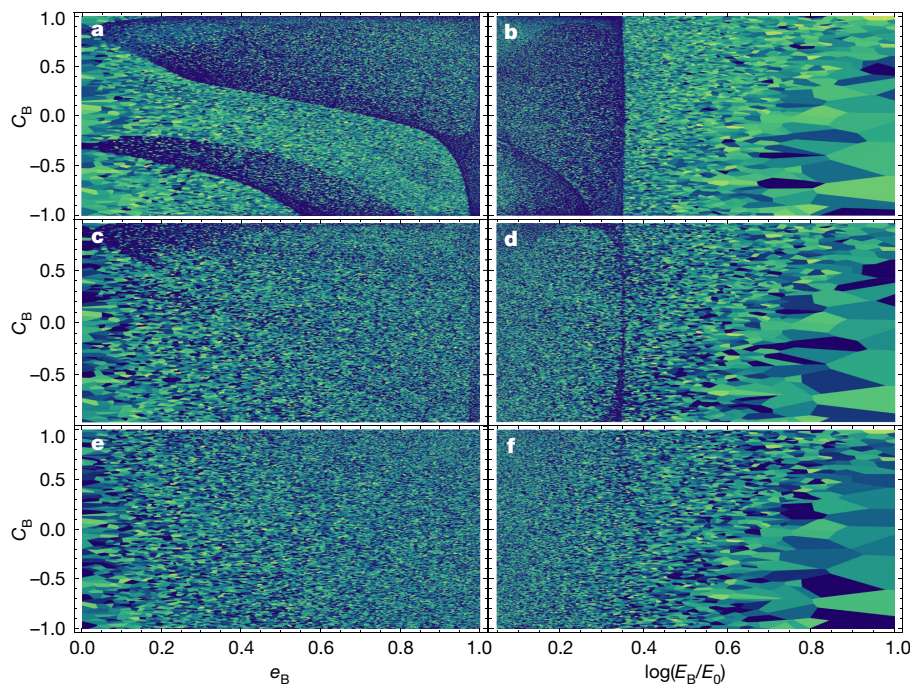


Fig. 2 | Topological maps of three-body scattering outcomes for run A. a–f. The total number of scrambles is colour-coded (smallest values of N_{scram} as dark blue, larger N_{scram} in green and yellow) with logarithmic scaling, as a function of survivor binary eccentricity e_B (**a**, **c**, **e**), energy E_B (**b**, **d**, **f**) and cosine inclination C_B . Shown are the cases $N_{\text{scram}} \geq 0$ (**a**, **b**), $N_{\text{scram}} \geq 1$ (**c**, **d**) and $N_{\text{scram}} \geq 2$ (**e**, **f**). Clouds of regularity obscure the underlying chaotic sea in **a**, **b**, but have dissipated in **e**, **f**, indicating that scrambles are the key dynamical mechanism responsible for ‘ergodizing’ the comparable-mass three-body problem.

parametrize L_0 in terms of the initial binary eccentricity e_0). However, many of our scattering experiments do not form resonant three-body systems, but instead resolve abruptly in a prompt exchange, where it is unlikely that the ergodic hypothesis can be applied. Metastable three-body systems generally exhibit intermittent chaos²³. Long periods of quasi-regular evolution occur during the non-terminal ejection of a

single star, but these are then interrupted by brief periods of intensely chaotic evolution when that star returns to the pericentre^{4,10}. We hypothesize that the degree of ergodicity in a subset of scattering experiments can be inferred from the number of scrambles, N_{scram} .

We illustrate the development of ergodicity in Fig. 2, which shows topological maps in outcome space. Whereas the full scattering ensemble

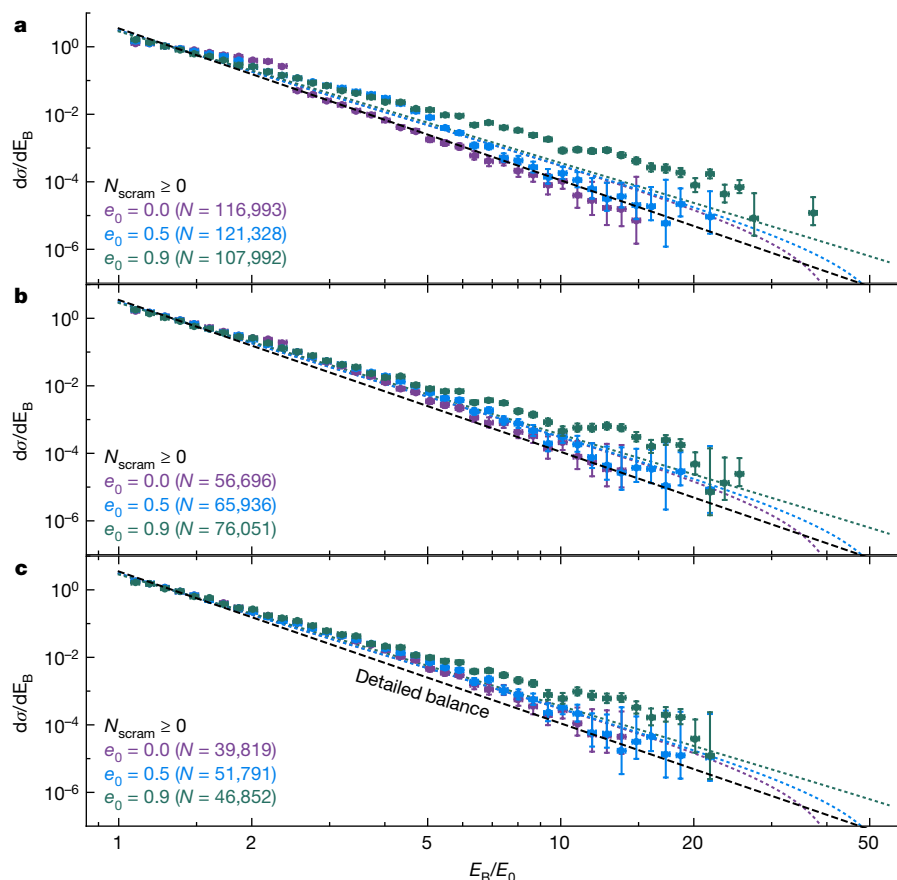


Fig. 3 | Marginal distribution of binary energy, $d\sigma/dE_B$, as a function of dimensionless energy, E_B/E_0 . The dotted lines are ergodic outcome distributions for ensembles with high (purple), medium (blue) and low (green) angular momentum. The data points are binned outcomes from numerical binary–single scattering ensembles ($N \approx 10^5$). Horizontal error bars show bin sizes and vertical error bars indicate 95% Poissonian confidence intervals. **a**, Full set of results from our numerical scattering experiments. **b**, Subset of results for $N_{\text{scram}} \geq 1$. **c**, Subset of results for $N_{\text{scram}} \geq 2$. Detailed balance (black dashed line) is never achieved.

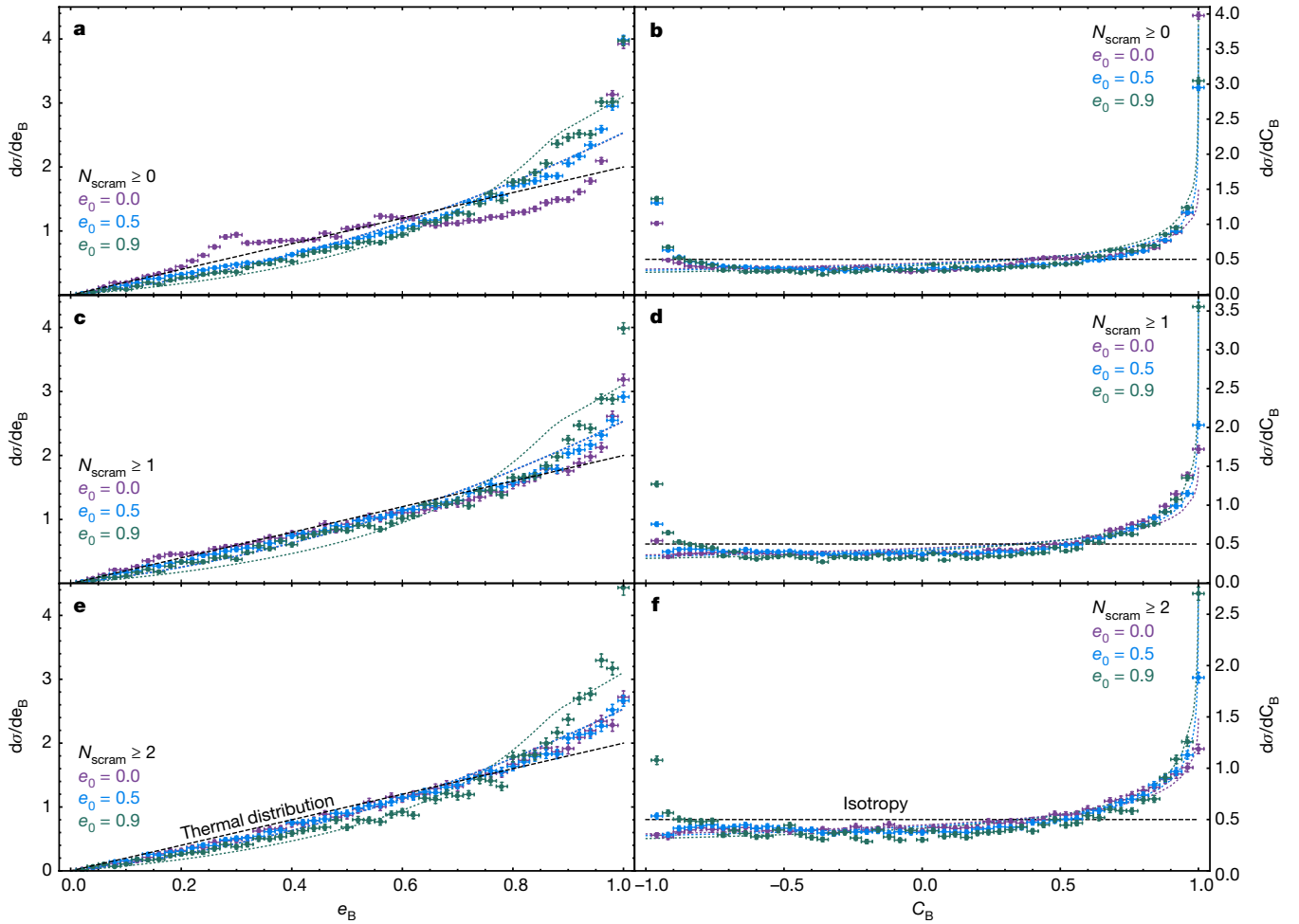


Fig. 4 | Marginal distributions of binary eccentricity and orientation. **a, c, e,** $d\sigma/de_B$ against eccentricity e_B . **b, d, f,** $d\sigma/dC_B$ against the cosine of the binary inclination, C_B . Line styles represent ergodic outcome distributions with the same ensemble angular momenta as in Fig. 3. The data points are binned outcomes from the same numerical scattering ensembles as in Fig. 3, with each row corresponding to the same cuts on N_{scram} . The eccentricity outcome

distributions are notably super-thermal (the thermal distribution, $d\sigma/de_B = 2e$, is shown as a black dashed line). The inclination distributions exhibit anisotropic bias towards prograde binaries aligned with L_0 (the isotropic distribution is shown with a black dashed line). Horizontal error bars show bin sizes and vertical error bars indicate 95% Poissonian confidence intervals.

has clear geometrical features indicative of prompt exchanges, these ‘clouds of regularity’ mostly (entirely) disappear if one considers the $\sim 50\%$ of integrations with $N_{\text{scram}} \geq 1$ ($N_{\text{scram}} \geq 2$). With this qualitative argument in mind, we use Figs. 3 and 4 to quantitatively compare the binned results of our scattering experiments to the marginal distributions predicted by the ergodic hypothesis. Horizontal error bars show bin sizes and vertical error bars indicate 95% Poissonian confidence intervals. All three of the marginal distributions that we examine ($d\sigma/dE_B$, $d\sigma/de_B$ and $d\sigma/dC_B$) exhibit reasonable (and sometimes very close) agreement between the ergodic theory of equation (2) and our numerical scattering experiments, provided that we examine resonant encounters ($N_{\text{scram}} \geq 2$). The marginal distributions for large- L_0 ensembles are very consistent with the numerical experiments. The agreement is slightly worse for our low- L_0 ensemble.

The agreement between ergodic theory and experiment is never exact, even in $N_{\text{scram}} \geq 2$ subsamples, and in most cases we see data that match analytic predictions to leading order but also exhibit some level of higher-order structure. The nature of these superimposed, second-order structures is not altogether clear, as two explanations seem plausible. First, these could represent islands of regularity in the initial conditions that we explore: regions of parameter space that do not fully forget their initial conditions despite undergoing multiple scrambles. Second, these could represent a failure in the

idealized escape criteria, $R(E_B, L_B)$, that we employ. We only consider very simple definitions of the strong-interaction region, the true shape of which is probably connected to the stability boundary of the triple²⁴. We defer an investigation of these two hypotheses to future work.

Non-hierarchical triples are common, if short-lived, in the astrophysical Universe²⁵. They are responsible for many interesting phenomena. For example, binary–single scattering events in dense star clusters produce blue stragglers^{26,27}, cataclysmic variables²⁸, X-ray binaries^{29,30} and even binary stellar-mass black holes¹¹. The lattermost of these scenarios may be responsible for most of the black-hole mergers seen by the LIGO experiment^{12,13}. Dynamical formation of these systems in a binary–single scattering is favoured when the surviving binary is drawn from the high- e_B tail of outcomes. It is therefore notable that (i) we find generic superthermality in the outcomes of comparable-mass scatterings (both from ergodic theory and numerical experiments) and (ii) that our formalism has identified the type of binary–single encounters that are predisposed to produce exotic binaries: low- L_0 scatterings. In the future, it may be possible to apply our formalism to estimate the properties of temporary binaries formed during long, but non-terminal, single-star ejections. High eccentricity binaries formed as ‘intermediate states’ of a three-body resonance may merge during the ejection owing to short-range dissipative

forces, leading to, for example, uniquely eccentric gravitational-wave signals¹⁴.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1833-8>.

- Agekyan, T. A. & Anosova, Z. P. A study of the dynamics of triple systems by means of statistical sampling. *Astron. Zh.* **44**, 1261 (1967).
- Šuvakov, M. & Dmitrašinović, V. Three classes of Newtonian three-body planar periodic orbits. *Phys. Rev. Lett.* **110**, 114301 (2013).
- Standish, E. M. The dynamical evolution of triple star systems. *Astron. Astrophys.* **21**, 185–191 (1972).
- Hut, P. & Bahcall, J. N. Binary–single star scattering. I – numerical experiments for equal masses. *Astrophys. J.* **268**, 319–341 (1983).
- Poincaré, H. *Les Méthodes Nouvelles de la Mécanique Céleste* (Gauthier-Villars et fils, 1892).
- Sundman, K. F. Mémoire sur le problème des trois corps. *Acta Math.* **36**, 105–179 (1913).
- Fermi, E. High energy nuclear events. *Prog. Theor. Phys.* **5**, 570–583 (1950).
- Monaghan, J. J. A statistical theory of the disruption of three-body systems – I. Low angular momentum. *Mon. Not. R. Astron. Soc.* **176**, 63–72 (1976).
- Valtonen, M., Mylläri, A., Orlov, V. & Rubinov, A. Dynamics of rotating triple systems: statistical escape theory versus numerical simulations. *Mon. Not. R. Astron. Soc.* **364**, 91–98 (2005).
- Heggie, D. C. Binary evolution in stellar dynamics. *Mon. Not. R. Astron. Soc.* **173**, 729–787 (1975).
- Portegies Zwart, S. F. & McMillan, S. L. W. Black hole mergers in the Universe. *Astrophys. J. Lett.* **528**, 17–20 (2000).
- Rodriguez, C. L., Chatterjee, S. & Rasio, F. A. Binary black hole mergers from globular clusters: masses, merger rates, and the impact of stellar evolution. *Phys. Rev. D* **93**, 084029 (2016).
- Hong, J. et al. Binary black hole mergers from globular clusters: the impact of globular cluster properties. *Mon. Not. R. Astron. Soc.* **480**, 5645–5656 (2018).
- Samsing, J., MacLeod, M. & Ramirez-Ruiz, E. The formation of eccentric compact binary inspirals and the role of gravitational wave emission in binary–single stellar encounters. *Astrophys. J.* **784**, 71 (2014).
- Rodriguez, C. L. et al. Post-Newtonian dynamics in dense star clusters: formation, masses, and merger rates of highly-eccentric black hole binaries. *Phys. Rev. D* **98**, 123005 (2018).
- Portegies Zwart, S. F. & Boekholt, T. C. N. Numerical verification of the microscopic time reversibility of Newton's equations of motion: fighting exponential divergence. *Commun. Nonlinear Sci. Numer. Simul.* **61**, 160–166 (2018).
- Hut, P. The topology of three-body scattering. *Astron. J.* **88**, 1549–1559 (1983).
- Samsing, J. & Ilan, T. Topology of black hole binary–single interactions. *Mon. Not. R. Astron. Soc.* **476**, 1548–1560 (2018).
- Bohr, N. Neutron capture and nuclear constitution. *Nature* **137**, 344–348 (1936).
- Monaghan, J. J. A statistical theory of the disruption of three-body systems – II. High angular momentum. *Mon. Not. R. Astron. Soc.* **177**, 583–594 (1976).
- Nash, P. E. & Monaghan, J. J. A statistical theory of the disruption of three-body systems – III. Three-dimensional motion. *Mon. Not. R. Astron. Soc.* **184**, 119–125 (1978).
- Geller, A. M., Leigh, N. W. C., Giersz, M., Kremer, K. & Rasio, F. A. In search of the thermal eccentricity distribution. *Astrophys. J.* **872**, 165 (2019).
- Pomeau, Y. & Manneville, P. Intermittent transition to turbulence in dissipative dynamical systems. *Commun. Math. Phys.* **74**, 189–197 (1980).
- Mardling, R. A. & Aarseth, S. J. Tidal interactions in star cluster simulations. *Mon. Not. R. Astron. Soc.* **321**, 398–420 (2001).
- Leigh, N. W. C. & Geller, A. M. The dynamical significance of triple star systems in star clusters. *Mon. Not. R. Astron. Soc.* **432**, 2474–2479 (2013).
- Leonard, P. J. T. & Fahlman, G. G. On the origin of the blue stragglers in the globular cluster NGC 5053. *Astron. J.* **102**, 994 (1991).
- Leigh, N., Sills, A. & Knigge, C. An analytic model for blue straggler formation in globular clusters. *Mon. Not. R. Astron. Soc.* **416**, 1410–1418 (2011).
- Ivanova, N. et al. Formation and evolution of compact binaries in globular clusters – I. Binaries with white dwarfs. *Mon. Not. R. Astron. Soc.* **372**, 1043–1059 (2006).
- Pooley, D. & Hut, P. Dynamical formation of close binaries in globular clusters: cataclysmic variables. *Astrophys. J. Lett.* **646**, 143–146 (2006).
- Ivanova, N., Heinke, C. O., Rasio, F. A., Belczynski, K. & Fregeau, J. M. Formation and evolution of compact binaries in globular clusters – II. Binaries with neutron stars. *Mon. Not. R. Astron. Soc.* **386**, 553–576 (2008).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgements We acknowledge discussions with D. Hoggie, P. Hut, R. Sari and S. Portegies-Zwart, as well as feedback from E. Michaely and O. C. Winter. N.C.S. received financial support from NASA, through Einstein Postdoctoral Fellowship Award number PF5-160145 and the NASA Astrophysics Theory Research Program (grant NNX17AK43G; Principal Investigator, B. Metzger). N.C.S. also thanks the Aspen Center for Physics for its hospitality during early stages of this work. N.W.C.L. acknowledges support by Fondecyt Iniciación grant number 11180005. We thank the Chinese Academy of Sciences for hosting us as we completed our efforts. We thank M. Valtonen and H. Karttunen, whose book on the three-body problem motivated much of this work.

Author contributions N.C.S. led the analytic work, to which N.W.C.L. contributed significantly. The FEWBODY simulations were performed by N.W.C.L. The comparison between the simulations and the analytic theory was performed jointly by the two authors.

Competing interests The authors declare no competing interests.

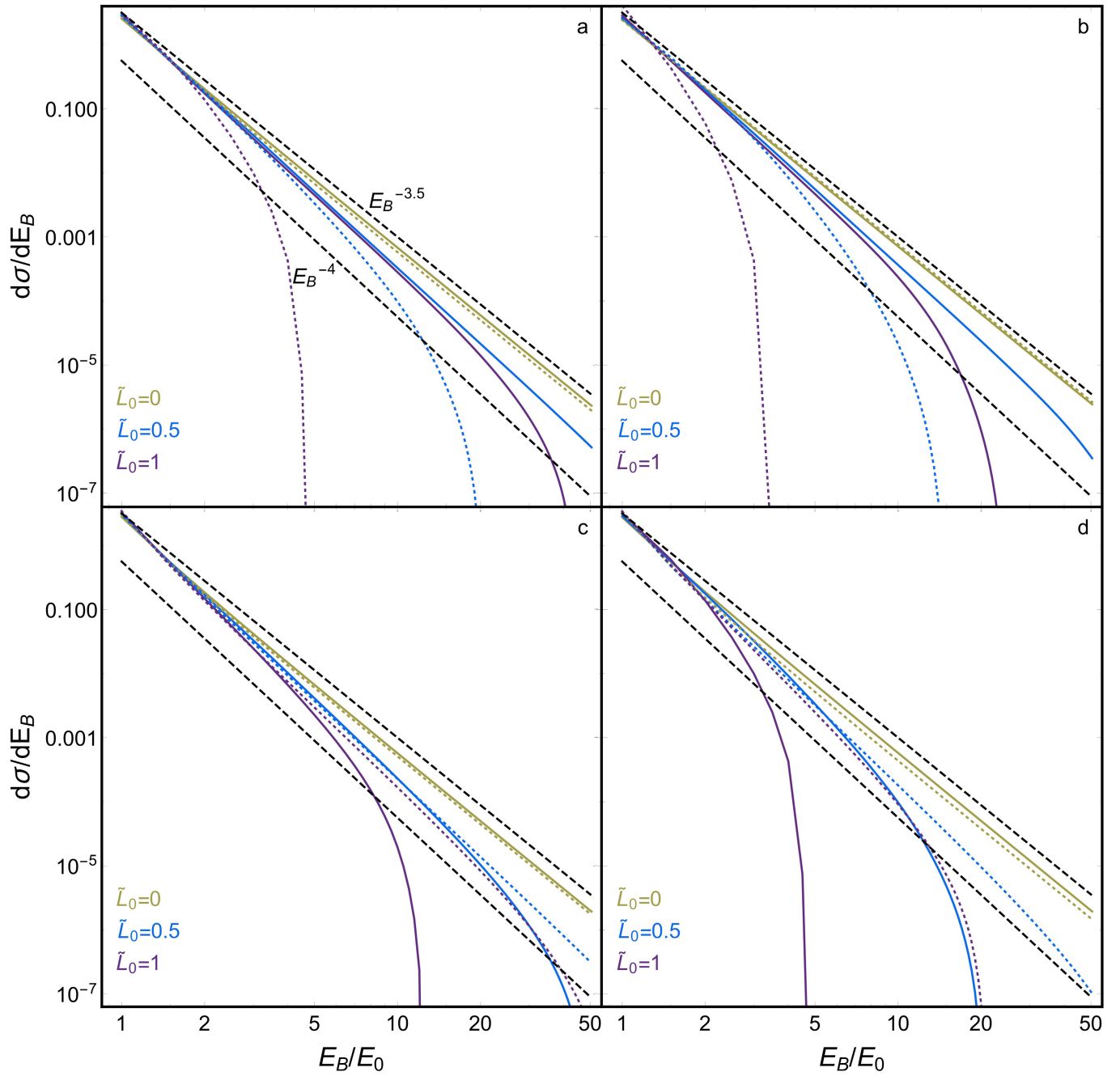
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1833-8>.

Correspondence and requests for materials should be addressed to N.C.S.

Peer review information *Nature* thanks Erez Michaely and Othon Cabo Winter for their contribution to the peer review of this work.

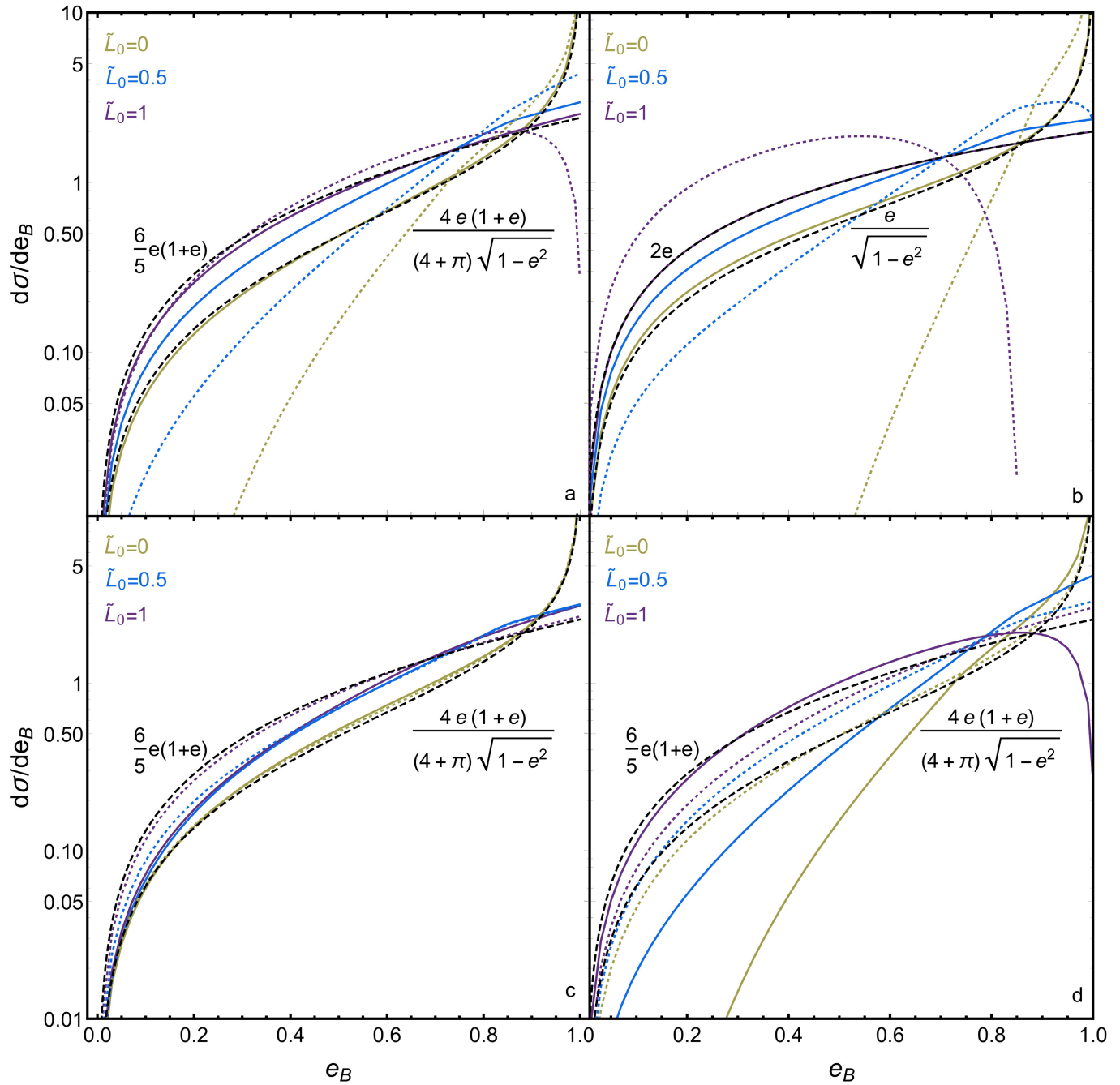
Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | Marginal distribution of binary energies, $d\sigma/dE_B$.

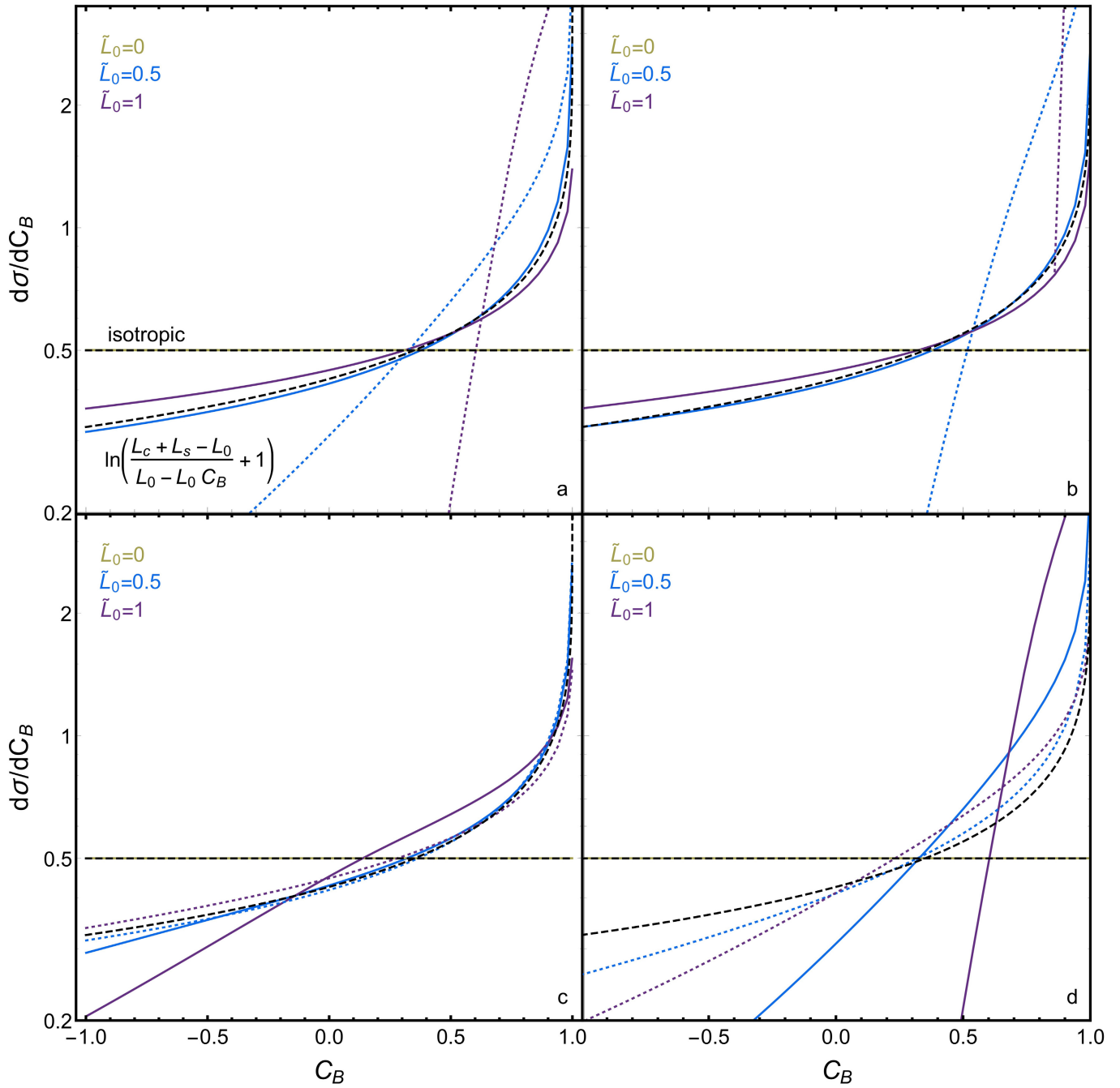
Colours show dimensionless angular momenta \tilde{L}_0 ; upper and lower black dashed lines are asymptotic power laws for $\tilde{L}_0 = 1$ and $\tilde{L}_0 \approx 1$, respectively. **a**, Ergodic outcome distributions using the ‘apocentric escape’ (AE) criterion; that is, assuming that disintegration of metastable triples occurs within a strong interaction region of size $R = \alpha a_b(1 + e_b)$. Here we take $\alpha = 2$. Solid lines represent equal-mass scattering ensembles ($m_a = m_b = m_s$) and dotted lines

extreme-mass-ratio ensembles ($m_a = m_b = 10m_s$). **b**, As in **a**, but for a ‘simple escape’ (SE) criterion, $R = \alpha a_b$. **c**, Intermediate-mass-ratio scattering ensembles ($m_a = m_b = 3m_s$). Solid lines correspond to $\alpha = 2$ and dotted lines to $\alpha = 5$. **d**, As in **c**, but for $m_a = m_b = 10m_s$. Note that \tilde{L}_0 is a dimensionless angular momentum normalized by the circular orbit angular momentum of a binary with energy E_0 and masses m_a and m_b .



Extended Data Fig. 2 | Marginal distribution of binary eccentricity, $d\sigma/de_B$. Line styles and assumptions are as in Extended Data Fig. 1, except for the upper and lower black dashed lines, which here show the $\tilde{L}_0 \approx 1$ and $\tilde{L}_0 \ll 1$ limits of the $d\sigma/de_B$ distribution, respectively (unlike for $d\sigma/dE_B$, these limits differ

significantly in the AE and SE regimes). In comparable-mass AE calculations, mildly super-thermal outcomes arise from geometric effects when $\tilde{L}_0 \approx 1$; by contrast, angular-momentum starvation produces extremely super-thermal outcomes when $\tilde{L}_0 \ll 1$. Small m_s values foreclose parts of e_B space, as $L_B \approx L_0$.



Extended Data Fig. 3 | Marginal distribution of binary orientation, $d\sigma/dC_B$. Assumptions and line styles are as in Extended Data Fig. 1, except that the black dashed lines show (i) an isotropic outcome configuration and (ii) an analytic approximation for $d\sigma/dC_B$, as labelled in **a** (for an equal-mass triple with $\tilde{L}_0 = 0.5$).

For $\tilde{L}_0 \ll 1$, surviving binaries are distributed isotropically (as symmetry dictates). Otherwise, binary orientations $C_B = \hat{\mathbf{L}}_B \cdot \hat{\mathbf{L}}_0$ are biased towards prograde outcomes. For extreme mass ratios and large \tilde{L}_0 , retrograde outcomes may be entirely prohibited.

Extended Data Table 1 | Numerical (binary–single) scattering ensembles used for comparison to analytic theory

Run	e_0	\tilde{L}_0	N_0	N_1	N_2
A	0.0	1.0	116,993	56,696	39,819
B	0.5	0.87	121,328	65,936	51,791
C	0.9	0.44	107,992	76,051	46,852

The first two columns show the initial binary eccentricity e_0 and the conserved dimensionless angular momentum \tilde{L}_0 in each simulated scattering run. The other columns show the number of runs with $N_{\text{scram}} \geq I, N_i$. Each run has initial impact parameter $b = 0$, isotropically distributed phase angles and particles of equal mass ($m_a = m_b = m_s$).

Atomic-scale imaging of a 27-nuclear-spin cluster using a quantum sensor

<https://doi.org/10.1038/s41586-019-1834-7>

Received: 10 May 2019

Accepted: 6 September 2019

Published online: 18 December 2019

M. H. Abobeih^{1,2}, J. Randall^{1,2}, C. E. Bradley^{1,2}, H. P. Bartling^{1,2}, M. A. Bakker^{1,2}, M. J. Degen^{1,2}, M. Markham³, D. J. Twitchen³ & T. H. Taminiau^{1,2*}

Nuclear magnetic resonance (NMR) is a powerful method for determining the structure of molecules and proteins¹. Whereas conventional NMR requires averaging over large ensembles, recent progress with single-spin quantum sensors^{2–9} has created the prospect of magnetic imaging of individual molecules^{10–13}. As an initial step towards this goal, isolated nuclear spins and spin pairs have been mapped^{14–21}. However, large clusters of interacting spins—such as those found in molecules—result in highly complex spectra. Imaging these complex systems is challenging because it requires high spectral resolution and efficient spatial reconstruction with sub-ångström precision. Here we realize such atomic-scale imaging using a single nitrogen vacancy centre as a quantum sensor, and demonstrate it on a model system of 27 coupled ¹³C nuclear spins in diamond. We present a multidimensional spectroscopy method that isolates individual nuclear–nuclear spin interactions with high spectral resolution (less than 80 millihertz) and high accuracy (2 millihertz). We show that these interactions encode the composition and inter-connectivity of the cluster, and develop methods to extract the three-dimensional structure of the cluster with sub-ångström resolution. Our results demonstrate a key capability towards magnetic imaging of individual molecules and other complex spin systems^{9–13}.

The nitrogen vacancy (NV) centre in diamond has emerged as a powerful quantum sensor^{2–13,22,23}. The NV electron spin provides long coherence times^{5,6,20} and high-contrast optical readout^{5,24,25}, enabling high sensitivity over a large range of temperatures^{5,6,20,25,26}. Pioneering experiments with near-surface NV centres have demonstrated spectroscopy of small ensembles of nuclear spins in nanoscale volumes^{2,3,5–8} and electron-spin-labelled proteins⁴. Furthermore, single-nuclear-spin sensitivity has been demonstrated, and isolated individual nuclear spins and spin pairs have been mapped^{14–21}. Together, these results have established the NV centre as a promising platform for magnetic imaging of complex spin systems and single molecules^{10–13}.

Here, we realize a key step towards that goal: the three-dimensional (3D) imaging of large nuclear-spin structures with atomic resolution. The main idea behind our method is to obtain structural information by accessing the couplings between individual nuclear spins. Our approach has three key elements: (1) realizing high spectral resolution so that small couplings can be accessed, (2) isolating such couplings from complex spectra, and (3) transforming the revealed connectivity into a 3D spatial structure with sub-ångström precision.

The basic elements of our experiment are illustrated in Fig. 1a. We consider a cluster of ¹³C nuclear spins in the vicinity of a single NV centre in diamond at 4 K. This cluster provides a model system for the magnetic imaging of single molecules and spin structures external to the diamond. Each ¹³C spin has a shifted frequency owing to the hyperfine interaction with the electron spin, resembling a chemical shift in traditional NMR^{1,27}. These shifts allow us to distinguish different nuclear spins in the cluster.

We use the NV electron spin as a sensor to probe nuclear–nuclear interactions (Fig. 1b). Inspired by NMR spectroscopy^{1,27}, we develop sequences that employ spin-echo double-resonance techniques to isolate and measure individual couplings with high spectral resolution. First, we polarize a nuclear ‘probe’ spin (frequency RF1) using recently developed quantum sensing sequences that can detect spins in any direction from the NV, enabling access to a large number of spins (see Methods)²⁸. Second, we let this probe spin evolve for a time t and apply N echo pulses that decouple it from the other spins and from environmental noise. Simultaneously, pulses on a ‘target’ spin in the cluster (frequency RF2) re-couple it to the probe spin, selecting the interaction between these two spins. Finally, a second sensing sequence detects the resulting polarization of the probe spin through high-contrast readout of the electron spin (see Methods), which enables fast data collection. This double-resonance sequence provides a high spectral resolution through a long nuclear phase accumulation time. Importantly, the resolution is not limited by the relatively short coherence time of the electron-spin sensor (see Methods)^{24,29}.

It is instructive to first consider the case without echo pulses ($N=0$). In such a Ramsey-type measurement^{24–26,29,30}, all couplings act simultaneously. This results in complex spectra that indicate the presence of multiple spins and many nuclear–nuclear spin interactions in the cluster (Fig. 1c). However, this one-dimensional (1D) measurement gives no direct information on the connectivity between spins. Additionally, the underlying structure of individual spins and couplings is obscured by the many frequencies (2^j for coupling to j spins) and by

¹QuTech, Delft University of Technology, Delft, The Netherlands. ²Kavli Institute of Nanoscience Delft, Delft University of Technology, Delft, The Netherlands. ³Element Six, Didcot, UK. *e-mail: T.H.Taminiau@TUDelft.nl

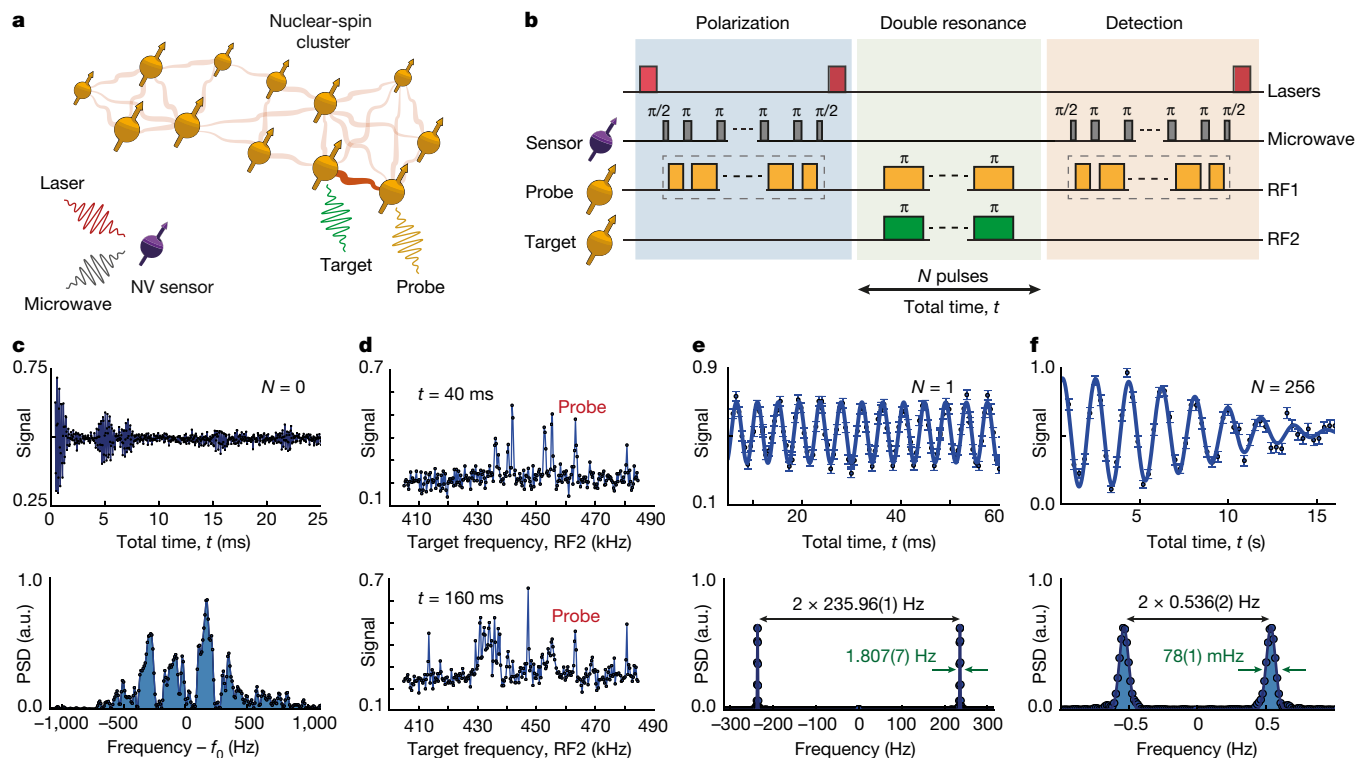


Fig. 1 | Basic concepts of the experiment. **a**, We consider an individual cluster of ^{13}C nuclear spins near a single NV centre in diamond. To obtain the 3D structure of the cluster we use the NV electron spin as a quantum sensor to measure nuclear–nuclear spin couplings. **b**, Experimental sequence. The NV sensor is used to polarize and detect the ‘probe’ spin(s) at frequency RF1 (see Methods). A double-resonance sequence of N echo pulses is applied simultaneously on the probe spin(s) (RF1) and the ‘target’ spin(s) (RF2), so that the coupling between these spins is selectively detected. See Extended Data Fig. 1 for the detailed sequence. **c**, A Ramsey signal ($N=0$) for a nuclear spin in the cluster (detuning $f_0 = 5$ kHz relative to RF1 = 455.37 kHz). Because all couplings are probed simultaneously, the power spectral density (PSD) yields a complex non-resolvable spectrum. See Extended Data Fig. 2 for more

examples. **d**, Double-resonance spectroscopy ($N=1$). Sweeping the target frequency (RF2) reveals all spins that couple to the probe spin(s). For larger t , more peaks appear as weaker couplings become visible. Here, RF1 = 463.27 kHz. **e**, Sweeping the evolution time t for a fixed RF1 and RF2 reveals the coupling strength between spins. This example reveals a 235.96(1) Hz coupling between two spins with a spectral resolution of 1.807(7) Hz FWHM. Here, RF1 = 463.27 kHz and RF2 = 455.37 kHz. **f**, An example with $N=256$ echo pulses showing a coherence time of 10.9(5) s, which enables selective measurements of sub-hertz couplings with high spectral resolution (78(1) mHz) and precision (2 mHz). Here, RF1 = 408.32 kHz and RF2 = 413.48 kHz. See Methods for the fit functions of all graphs. Error bars are one standard deviation. a.u., arbitrary units.

the low spectral resolution of >30 Hz full-width at half-maximum (FWHM), which is set by the dipolar-broadened linewidth of the nuclear spins and is inversely proportional to the dephasing time, T_2^* .

By contrast, our double-resonance sequence enables couplings between specific spins to be isolated and measured with high resolution. We first scan the target frequency RF2 for a fixed probe frequency RF1 (Fig. 1d). This reveals the spectral positions of nuclear spins coupled to the probe spin. We then sweep the evolution time t and apply a Fourier transform to the signal to quantify the coupling strengths (Fig. 1e). For a single pulse ($N=1$), the nuclear-spin coherence time is $T_2 = 0.58(2)$ s (all given uncertainties are one standard deviation), yielding a spectral resolution of 1.807(7) Hz and a centre frequency accuracy of 10 mHz. The spectral resolution is set by the coherence of the sample spins and can be further enhanced by applying more echo pulses. For $N=256$, a resolution of 78(1) mHz and an accuracy of 2 mHz are obtained, making it possible to detect sub-hertz interactions (Fig. 1f). The obtained resolution is improved by a factor about 10^3 compared with Ramsey-type spectroscopy on the same type of sample (Fig. 1c)^{18–21,24,26,29} and is an order of magnitude higher than that achieved in previous experiments on other spin samples^{6–8,25,30,31}.

To characterize the entire cluster, we perform 3D spectroscopy by varying the probe frequency RF1, the target frequency RF2 and the evolution time t . The combinations of RF1 and RF2 reveal the spectral positions of the spins in the cluster. The coupling between spins is

retrieved from the Fourier transform along the time dimension t . This yields a 3D dataset that in principle encodes the composition and connectivity of the spin cluster (Fig. 2).

In general, multiple spins can have (near-) identical precession frequencies. This has two consequences. First, the echo pulses will invert these spins simultaneously, so that multiple couplings are probed at the same time. Figure 3a shows an example with one probe spin and three target spins. This example illustrates that, although the resulting spectra are more complex, the high spectral resolution of our method enables retrieval of the underlying nuclear–nuclear couplings, even when several spins overlap spectrally.

Second, to determine the number of spins in the cluster and to assign the measured couplings to them, we need to resolve the ambiguity introduced by the fact that multiple spins can overlap spectrally. For example, the observation of a coupling between frequencies $\{f_a, f_b\}$ and a coupling between frequencies $\{f_b, f_c\}$ is by itself not enough to determine if there are one or two spins with frequency f_b . Our method resolves such ambiguities by extracting an over-determined dataset with many couplings that together constrain the problem. This enables individual spins to be uniquely identified from their connections to the rest of the cluster (see Fig. 3b for an example).

Transforming the 3D spectra into a spatial structure requires a precise relation between the measured couplings and the relative positions

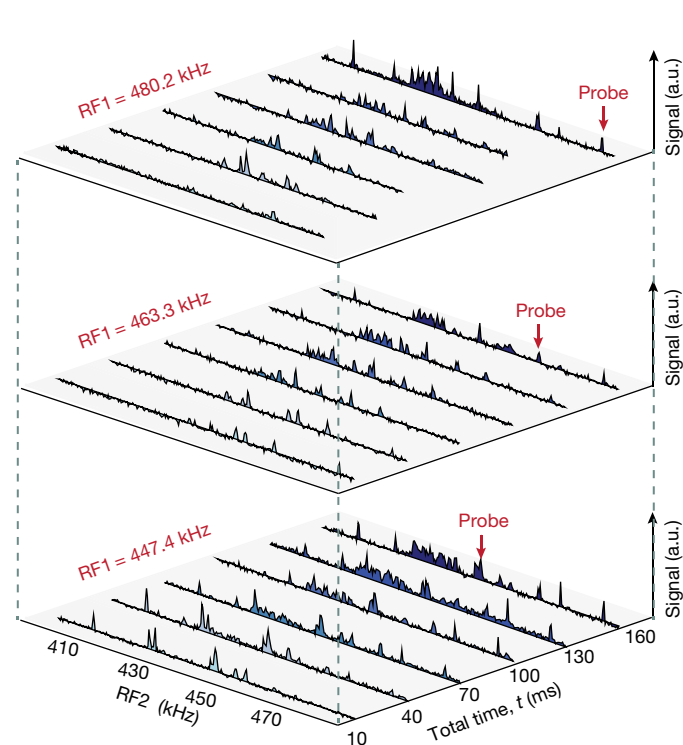


Fig. 2 | 3D spectroscopy. By varying the probe frequency RF1, the target frequency RF2 and the evolution time t , we obtain a 3D dataset that encodes the composition of the spins in the cluster and their couplings. The observation of a signal at {RF1, RF2} indicates the presence of one or more spins at both frequencies and a coupling between them. The Fourier transform along the time dimension t reveals the spin–spin coupling strengths. Examples for three different RF1 values are shown.

of the spins. A complication is that the presence of electronic spins can modify the nuclear couplings³², causing the measured value to deviate from a basic dipole–dipole coupling. We use perturbation theory to derive a set of many-body corrections that depend on the electron–nuclear and nuclear–nuclear couplings and on the magnetic-field direction (see Methods). For the type of cluster considered here, the corrections could be non-negligible. However, the signs of the leading terms depend on the electron spin state. By averaging the measured couplings for the $m_s = +1$ and $m_s = -1$ states (m_s , spin projection), the deviations are strongly reduced. Together with the use of a novel method to align the magnetic field along the NV axis to within 0.07° (see Methods), this enables us to approximate the nuclear–nuclear couplings as dipolar.

Finally, we determine the structure of the spin cluster. Figure 4a summarizes all extracted couplings. We identify $M = 27$ nuclear spins and retrieve 171 pairwise couplings out of the total of $M(M-1)/2 = 351$ couplings. The structure of the cluster is fully described by $3M - 4 = 77$ spatial coordinates (see Methods), so the problem is over-determined. However, owing to the large number of parameters and local minima, a direct least-squares minimization¹⁰ is challenging. Instead, we build the structure sequentially by progressively adding spins while keeping track of all possible structures that match the measured couplings within a certain tolerance.

We use two different methods. The first method constrains the spin coordinates to the diamond lattice. The second method discretizes space in a general cubic lattice, with voxel spacing as low as 5×10^{-3} nm (about 1/70 of the lattice constant; see Methods). Although this second method is more computationally intensive, it uses minimum a priori knowledge and can be applied to arbitrary spin systems. We run these analyses in parallel with the measurements, so that sets of the most

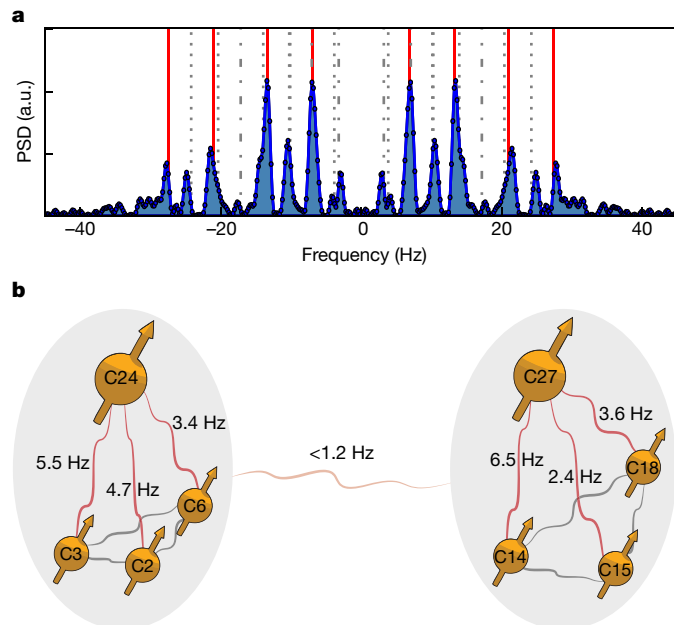


Fig. 3 | Spectrally overlapping spins. **a**, Retrieving couplings when multiple spins are re-coupled simultaneously. Example in which the echo pulses invert three target spins (quadruple resonance). The PSD reveals a complex, yet resolvable, spectrum. Red lines indicate the eight frequencies $f = \pm f_1 \pm f_2 \pm f_3$, where $f_1 = 17.17(2)$ Hz, $f_2 = 7.05(3)$ Hz and $f_3 = 3.21(4)$ Hz are the extracted couplings of the probe spin to three target spins. Grey dotted (dot-dashed) lines mark additional frequency components that appear owing to failures to invert one (two) of the target spins (see Supplementary Fig. 1 for detailed analysis). **b**, Overcoming ambiguity in identifying spins and assigning couplings. Shown is an example from the data. Spins C2, C3, C6, C14, C15 and C18 all yield a coupling signal to the same RF2 frequency. Because the couplings between these six spins reveal that they are part of two spatially separated sub-clusters, it follows that the signals at RF2 must originate from two distinct spins (C24 and C27).

promising spin assignments and structures are regularly created. These yield predictions for which unmeasured couplings (combinations of RF1 and RF2) are required to choose between different assignments and structures, which we use to guide the experiments and reduce the total measurement time (see Methods).

Figure 4b shows the structure obtained for the 27 spins using the diamond lattice. The blue lines show the strongest couplings (>3 Hz) and visualize the inter-connectivity of the cluster. The cubic-lattice method yields a nearly identical structure (see Methods); the average distance between the spin positions for the two solutions is 0.58 Å, a fraction of the bond length of about 1.54 Å. As a final step, we use these structures as inputs for least-squares minimization, where the x , y , z coordinates are allowed to relax to any value. The solution obtained lies close to the initial guess, with an average distance of 0.46 Å. The uncertainties for the spatial coordinates (δx , δy , δz) are below one diamond bond length for all 27 spins (Fig. 4c, d), indicating atomic-scale imaging of the complete 27-spin cluster.

Additionally, we determine the position of the NV sensor relative to the cluster. Although not required to reconstruct the cluster, this provides a control experiment. We measure the coupling of the ^{14}N nuclear spin to 12 of the ^{13}C spins (Extended Data Fig. 4). This unambiguously determines the location of both the ^{14}N atom and the vacancy (fit uncertainties <0.3 Å). We can now compare the electron– ^{13}C hyperfine couplings to previous density functional theory (DFT) calculations for 5 of our spins³³. All 5 couplings agree with the DFT calculations (Extended Data Fig. 4), providing an independent corroboration of the extracted structure, as well as a direct test of the DFT calculations.

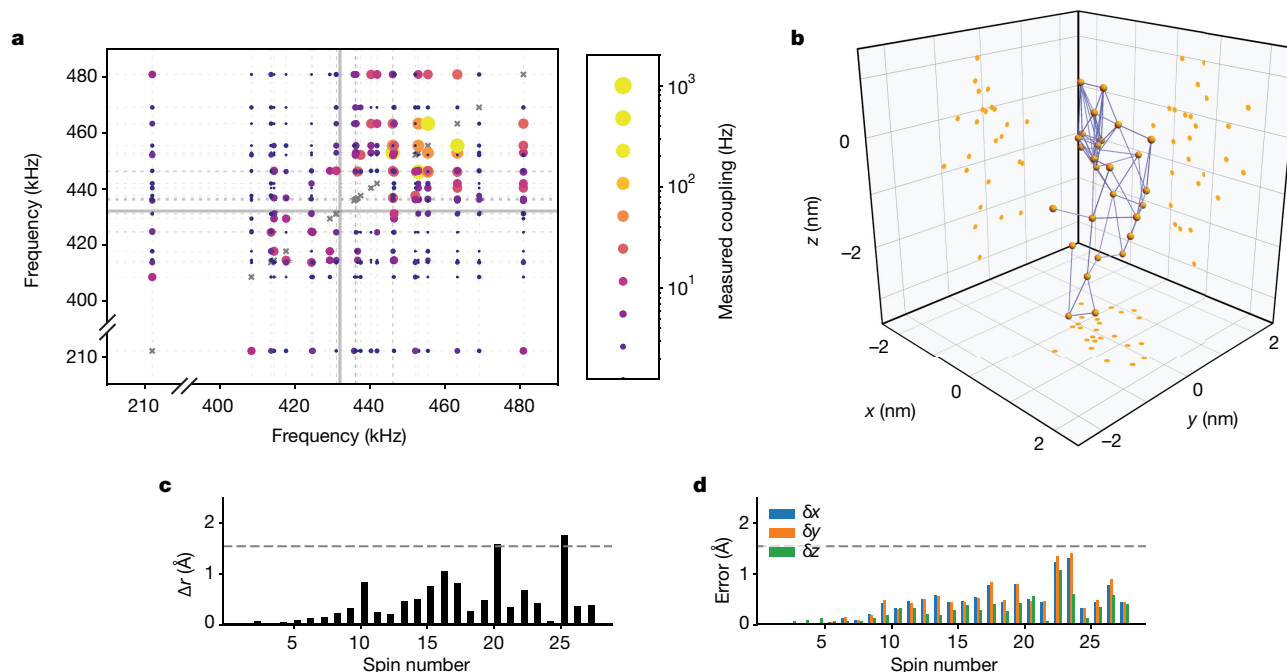


Fig. 4 | Atomic-scale imaging of the 27-nuclear-spin cluster. **a**, Two-dimensional (2D) plot summarizing all couplings between the 27 spins identified by 3D spectroscopy (Fig. 2), including identification of spins with overlapping frequencies. The size and colour of each point indicate the strength of the measured coupling, averaged over the electron $m_s = +1$ and $m_s = -1$ states. Dashed grey lines indicate the nuclear spin frequencies ($m_s = -1$ state). Solid grey lines indicate the bare ^{13}C Larmor frequency. Total measurement time, ~400 h. See Supplementary Tables 2–4 for numerical values and uncertainties. **b**, 3D structure of the nuclear spins, obtained using the diamond-lattice method (see text). Blue lines indicate couplings greater

than 3 Hz and illustrate the connectivity of the cluster. See Supplementary Video 1 for a visualization and Extended Data Fig. 3 for zoom-ins of strongly coupled subclusters. **c**, Distance Δr between the obtained spin positions from the diamond-lattice method (see text) and from a least-squares optimization. Deviations are generally below one diamond bond length (~1.54 Å, dashed line). **d**, The uncertainties for the 77 spatial coordinates of the cluster from a least-squares optimization are smaller than the bond length, indicating atomic-scale resolution. See Supplementary Figs. 4–6 and Supplementary Table 5 for in-depth comparisons between the structures and uncertainties obtained with different methods.

Looking beyond quantum sensing, this precise microscopic characterization of the NV environment provides new opportunities for improved control of quantum bits for quantum information^{20,24,28,31,32} and for investigating many-body physics in coupled spin systems.

In our method, the NV sensor spin is exclusively used to create and detect polarization (Fig. 1b). Therefore, the two main requirements for the sensor spin are (1) a high-contrast readout to keep measurement times manageable and (2) that it does not limit the spectral resolution by disturbing the evolution of the nuclear spins through relaxation^{25,30,31}. We satisfy these requirements by working at 4 K, so that the electron relaxation is negligible ($T_1 = 3.6(3) \times 10^3$ s)²⁰, and high-fidelity readout is obtained through resonant optical excitation (see Methods). Recent experiments have demonstrated both these requirements up to room temperature^{5,25,26,30,31}. The electron-spin relaxation—milliseconds at room temperature—can be decoupled from the sample spins through laser illumination^{30,31} or sequential weak measurements^{25,26}. High-contrast readout has been demonstrated by using a nuclear spin as a memory that can be read out repeatedly^{5,30}. Nuclear spins themselves are well isolated from temperature effects³¹. Therefore, when combined with those methods, the ideas presented here could be extended to ambient conditions.

In conclusion, we have developed and demonstrated 3D atomic-scale imaging of large clusters of nuclear spins using a single-spin quantum sensor. Our approach is compatible with room-temperature operation^{25,26,30,31} and could be extended to larger structures, as the number of required measurements scales linearly with the number of spins. Future improvements in the data acquisition and the computation of 3D structures can further reduce time requirements. In particular, recent methods to polarize and measure nuclear spins are expected to improve sensitivity^{25,26}, especially for samples with weak couplings to

the NV sensor. Optimized sampling of the measurements and adaptive algorithms based on real-time structure analysis can further reduce the total number of required measurements. Therefore, when combined with recent advances in nanoscale NMR with near-surface NV centres^{2–8}, our results provide a path towards magnetic imaging of individual molecules and complex spin structures external to diamond^{10–13}.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1834-7>.

- Rule, G. S. & Hitchens, T. K. *Fundamentals of Protein NMR Spectroscopy* Vol. 5 (Springer Science & Business Media, 2006).
- Mamin, H. et al. Nanoscale nuclear magnetic resonance with a nitrogen-vacancy spin sensor. *Science* **339**, 557–560 (2013).
- Staudacher, T. et al. Nuclear magnetic resonance spectroscopy on a (5-nanometer)³ sample volume. *Science* **339**, 561–563 (2013).
- Shi, F. et al. Single-protein spin resonance spectroscopy under ambient conditions. *Science* **347**, 1135–1138 (2015).
- Lovchinsky, I. et al. Nuclear magnetic resonance detection and spectroscopy of single proteins using quantum logic. *Science* **351**, 836–841 (2016).
- Aslam, N. et al. Nanoscale nuclear magnetic resonance with chemical resolution. *Science* **357**, 67–71 (2017).
- Glenn, D. R. et al. High-resolution magnetic resonance spectroscopy using a solid-state spin sensor. *Nature* **555**, 351–354 (2018).
- Smits, J. et al. Two-dimensional nuclear magnetic resonance spectroscopy with a microfluidic diamond quantum sensor. *Sci. Adv.* **5**, eaaw7895 (2019).
- Lovchinsky, I. et al. Magnetic resonance spectroscopy of an atomically thin material using a single-spin qubit. *Science* **355**, 503–507 (2017).

10. Ajoy, A., Bissbort, U., Lukin, M. D., Walsworth, R. L. & Cappellaro, P. Atomic-scale nuclear spin imaging using quantum-assisted sensors in diamond. *Phys. Rev. X* **5**, 011001 (2015).
11. Kost, M., Cai, J. & Plenio, M. B. Resolving single molecule structures with nitrogen-vacancy centers in diamond. *Sci. Rep.* **5**, 11007 (2015).
12. Perunicic, V., Hill, C., Hall, L. & Hollenberg, L. A quantum spin-probe molecular microscope. *Nat. Commun.* **7**, 12667 (2016).
13. Wang, Z.-Y., Haase, J. F., Casanova, J. & Plenio, M. B. Positioning nuclear spins in interacting clusters for quantum technologies and bioimaging. *Phys. Rev. B* **93**, 174104 (2016).
14. Sushkov, A. et al. Magnetic resonance detection of individual proton spins using quantum reporters. *Phys. Rev. Lett.* **113**, 197601 (2014).
15. Müller, C. et al. Nuclear magnetic resonance spectroscopy with single spin sensitivity. *Nat. Commun.* **5**, 4703 (2014).
16. Shi, F. et al. Sensing and atomic-scale structure analysis of single nuclear-spin clusters in diamond. *Nat. Phys.* **10**, 21–25 (2014).
17. Zopes, J. et al. Three-dimensional localization spectroscopy of individual nuclear spins with sub-angstrom resolution. *Nat. Commun.* **9**, 4678 (2018).
18. Zopes, J., Herb, K., Cuijia, K. S. & Degen, C. L. Three-dimensional nuclear spin positioning using coherent radio-frequency control. *Phys. Rev. Lett.* **121**, 170801 (2018).
19. Sasaki, K., Itoh, K. M. & Abe, E. Determination of the position of a single nuclear spin from free nuclear precessions detected by a solid-state quantum sensor. *Phys. Rev. B* **98**, 121405 (2018).
20. Abobeih, M. H. et al. One-second coherence for a single electron spin coupled to a multi-qubit nuclear-spin environment. *Nat. Commun.* **9**, 2552 (2018).
21. Yang, Z. et al. Structural analysis of nuclear spin clusters via two-dimensional nanoscale nuclear magnetic resonance spectroscopy. Preprint at <https://arxiv.org/abs/1902.05676v2> (2019).
22. Rosenfeld, E. L., Pham, L. M., Lukin, M. D. & Walsworth, R. L. Sensing coherent dynamics of electronic spin clusters in solids. *Phys. Rev. Lett.* **120**, 243604 (2018).
23. Knowles, H. S., Kara, D. M. & Atatüre, M. Demonstration of a coherent electronic spin cluster in diamond. *Phys. Rev. Lett.* **117**, 100802 (2016).
24. Cramer, J. et al. Repeated quantum error correction on a continuously encoded qubit by real-time feedback. *Nat. Commun.* **7**, 11526 (2016).
25. Pfender, M. et al. High-resolution spectroscopy of single nuclear spins via sequential weak measurements. *Nat. Commun.* **10**, 594 (2019).
26. Cuijia, K., Boss, J., Herb, K., Zopes, J. & Degen, C. Tracking the precession of single nuclear spins by weak measurements. *Nature* **571**, 230–233 (2019).
27. Slichter, C. *Principles of Magnetic Resonance* (Springer, 1996).
28. Bradley, C. E. et al. A ten-qubit solid-state spin register with quantum memory up to one minute. *Phys. Rev. X* **9**, 031045 (2019).
29. Laraoui, A. et al. High-resolution correlation spectroscopy of ^{13}C spins near a nitrogen-vacancy centre in diamond. *Nat. Commun.* **4**, 1651 (2013).
30. Pfender, M. et al. Nonvolatile nuclear spin memory enables sensor-unlimited nanoscale spectroscopy of small spin clusters. *Nat. Commun.* **8**, 834 (2017).
31. Maurer, P. C. et al. Room-temperature quantum bit memory exceeding one second. *Science* **336**, 1283–1286 (2012).
32. Dutt, M. G. et al. Quantum register based on individual electronic and nuclear spin qubits in diamond. *Science* **316**, 1312–1316 (2007).
33. Nizovtsev, A. P. et al. Non-flipping ^{13}C spins near an NV center in diamond: hyperfine and spatial characteristics by density functional theory simulation of the $\text{C}_{510}[\text{NV}]\text{H}_{252}$ cluster. *New J. Phys.* **20**, 023022 (2018).

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Sample and NV centre sensor

We use a naturally occurring NV centre in a diamond grown homoepitaxially by chemical vapour deposition with a 1.1% (natural) abundance of ^{13}C and a $\langle 111 \rangle$ crystal orientation (Element Six). The NV is placed in a solid-immersion lens to enhance the photon-collection efficiency³⁴. The NV centre was selected for the absence of ^{13}C spins with hyperfine couplings >500 kHz. The NV electron-spin dephasing time is $T_2^* = 4.9(2)$ μs and the echo coherence time is $T_2 = 1.182(5)$ ms. We work at 4 K, so that the electron relaxation is negligible ($T_1 = 3.6(3) \times 10^3$ s)²⁰ and use high-fidelity readout through resonant optical excitation (average fidelity $F = 94.5\%$)³⁴.

The observed nuclear-spin dephasing times range from $T_2^* = 3$ ms to 17 ms, corresponding to an inhomogeneous linewidth of about 30–150 Hz. Owing to the frequency difference between nuclear spins in $m_s = \pm 1$ (Supplementary Table 1), spin diffusion is strongly suppressed and the longitudinal relaxation of the nuclear spins is $T_1 > 6$ min (ref. ²⁸).

Magnetic-field alignment

A magnetic field of ~ 403 G is applied using a room-temperature permanent magnet, which is installed on an XYZ translation stage to control the strength and the direction of the magnetic field. Our methods use echo pulses and are therefore robust against slow fluctuations in the magnetic-field strength. Although magnetic-field drift has a negligible effect on the measured nuclear–nuclear couplings, we stabilize the magnetic field to <3 mG using temperature stabilization of the magnet and an automatic re-calibration procedure (every few hours).

We align the magnetic field along the NV axis to avoid electron-mediated shifts that cause the measured couplings to deviate from nuclear–nuclear dipolar coupling (see Supplementary Information section III). We use a ‘thermal’ echo sequence, which was previously used to measure the temperature³⁵ (see Extended Data Fig. 5). In this sequence, the electron evolves half of the time in a superposition of the states $m_s = 0$ and $m_s = -1$ and the other half in a superposition of $m_s = 0$ and $m_s = +1$. Because the energies of the states $m_s = \pm 1$ are shifted by equal and opposite amounts by Hamiltonian terms proportional to the spin operator S_z , the effects of such terms are cancelled. However, Hamiltonian terms that shift the energies of $m_s = \pm 1$ in the same way (such as the magnetic field perpendicular to z) do not cancel. Therefore, the sequence decouples the main source of noise (the magnetic-field fluctuations along z from the surrounding spin bath) while remaining sensitive to shifts caused by a non-zero magnetic field in the x, y directions. This sequence extends the sensing time from the dephasing time $T_2^* \approx 5$ μs to the echo coherence time $T_2 \approx 1$ ms, resulting in an uncertainty of 0.07° in the alignment (Extended Data Fig. 5).

Quantum sensing sequences

We employ two different sensing sequences (see the polarization and detection blocks in Fig. 1b). Sequence A consists of dynamical decoupling sequences of N' equally spaced π pulses on the electron spin of the form^{36–38} $(\tau_r - \pi - \tau_r)^{N'}$. This sequence is sensitive only to nuclear spins with a substantial electron–nuclear hyperfine component perpendicular to the applied magnetic field³⁶. The inter-pulse spacing $2\tau_r$ determines the spin frequency that is being probed.

Sequence B is a recently developed method, described in detail in Bradley et al.²⁸, that interleaves the dynamical decoupling sequence with radiofrequency (RF) pulses. This method enables the detection of spins with a weak or negligible perpendicular hyperfine component^{28,30}. For this sequence, the frequency of the RF pulse sets the targeted spin frequency, whereas τ_r can be chosen freely²⁸. Importantly, the amplitudes and phases of the RF pulses are set so that together they build up to the desired evolution²⁸. The added RF field imprints a deterministic phase on the electron-spin sensor²⁸, which we compensate by calibrating the phase of the electron $\pi/2$ pulses.

Electron–nuclear spectroscopy

As a starting point, we use the electron spin as a sensor to roughly characterize some of the nuclear spins in the cluster. We perform spectroscopy by sweeping the interpulse delay τ_r in sequence A (see, for example, Abobeih et al.²⁰) and the RF frequency for sequence B (ref. ²⁸). This identifies the frequency range in which spins are present in the cluster and provides the parameters to polarize and detect several spins²⁴. We note that the resolution of this spectroscopy technique is limited by the electron spin T_2 and the nuclear spin T_2^* .

Nuclear–nuclear double-resonance spectroscopy

The sequence for the double-resonance experiments is shown in Fig. 1b and Extended Data Fig. 1. To polarize and detect the probe spin, we use either sequence A (without the RF1 pulses in the dashed box) or sequence B (with the RF1 pulses), depending on whether the perpendicular hyperfine coupling to the electron spin is sufficiently large or not. For sequence A, we set the interpulse delay as $\tau_r = (2k - 1)\pi/(\omega_0 + \text{RF1})$, with k an integer and ω_0 the ^{13}C Larmor frequency for the $m_s = 0$ electron state, and calibrate the number of pulses N' to maximize the signal³⁶. For sequence B we calibrate the RF power to maximize the signal.

We create nuclear polarization by projective measurements²⁴. First the electron is prepared in a superposition state through resonant excitation³⁴ and a $\pi/2$ pulse. Second, the sensing sequence correlates the phase of the electron with the nuclear spin state. Finally, the electron is read out so that the nuclear spin is projected to a polarized state²⁴. To enhance the signal-to-noise ratio and to ensure that the electron measurement does not disturb the nuclear-spin evolution, we perform the double-resonance sequence only if a photon is detected during the electron readout²⁴. The resulting signal contrast for different spins varies from 20% to 96%.

Because the correlation data are read out and stored in the electronics, the ultimate limit for the spectral resolution of our method—that is, when applied on hypothetical signals with infinitesimal spectral width—is set by the precision of the 10-MHz reference clock used for the timing of the waveform generator^{7,39,40}. For the double-resonance sequence, the phases of the RF1 echo pulses are calibrated so that their phase difference is 0 or $\pi/2$ with respect to the polarization axis, which is determined by the direction of the hyperfine interaction^{18,19,41}. For the target spins, the phase of the RF2 pulse does not affect the signal and is set arbitrarily.

To mitigate pulse errors, we alternate the phases of the pulses following the XY8 scheme⁴², for both the electron and nuclear spins. For the electron spin, we use Hermite pulse envelopes⁴³ with a Rabi frequency of ~ 14 MHz to obtain effective microwave pulses without initialization of the intrinsic ^{14}N nuclear spin. The nuclear-spin Rabi frequencies are in the range 0.3–0.7 kHz.

Data analysis

We extract the spin–spin couplings f and their uncertainties from fitting the time-domain double-resonance signals (for example, Fig. 1e, f, top) with $S = a + Ae^{-(t/T_2)^n} \cos(2\pi ft + \phi)$, where a, A, ϕ and T_2 are fit parameters that account for the offset, amplitude, phase and coherence time of the signal, respectively. The PSD is obtained from a Fourier transform of the time-domain signal with zero filling¹ and the d.c. component filtered out (for example, Fig. 1e, f, bottom). The spectral resolution (FWHM) is obtained from a Gaussian fit of the PSD. Alternatively, we can define the spectral resolution (FWHM) directly from the time-domain signal as $2\sqrt{\ln 2}/(\pi T_2)$. This yields a spectral resolution of 0.91(3) Hz for Fig. 1e. For the spin in Fig. 1f, using $N = 1$ yields a spectral resolution of 0.8(1) Hz and using $N = 256$ yields 49(2) mHz. We note that no saturation is observed in the improvement of the spectral resolution with the number of pulses. Therefore, with more pulses (and longer measurement times) higher spectral resolutions and more precise measurements are feasible.

Electron-mediated interactions

We calculate corrections to the nuclear–nuclear couplings caused by the presence of the electron spin using perturbation theory up to second order. The effect of other nuclear spins on nuclear–nuclear couplings was found by numerical simulations to be negligible (of the order of millihertz). In contrast to previous results for strong electron–nuclear couplings^{32,44}, here many-body interactions due to non-secular nuclear–nuclear couplings must be taken into account. The resulting frequency in a double-resonance experiment is of the form (see Supplementary Information section III)

$$f_{\text{DR}}(m_s = \pm 1) \approx \frac{1}{4\pi} |C + \Delta\lambda_1(m_s) + \Delta\lambda_2(m_s) + \Delta\lambda_3(m_s)| \quad (1)$$

where C is the parallel (zz) component of the dipole–dipole interaction between the nuclear spins and $\Delta\lambda_i$ are correction terms accounting for the presence of the electron spin. See Supplementary Information for the full analysis of all terms.

The dominant correction for our parameter regime is $\Delta\lambda_2$, which depends on both the electron–nuclear and nuclear–nuclear interactions. We make a Taylor expansion up to first order in $A_{zz}^{(j)}/(\gamma_c B_z)$, where $A_{zz}^{(j)}$ is the parallel electron–nuclear hyperfine coupling for spin j , γ_c is the nuclear gyromagnetic ratio and B_z is the component of the magnetic field along the NV axis. This yields an expression of the form $\Delta\lambda_2(m_s) \approx m_s \Delta\lambda_2^{(0)} + \Delta\lambda_2^{(1)}$, where the leading (zeroth-order) correction $m_s \Delta\lambda_2^{(0)}$ is given by

$$\Delta\lambda_2^{(0)} = \frac{(A_{zx}^{(1)} + A_{zx}^{(2)})C_{zx} + (A_{zy}^{(1)} + A_{zy}^{(2)})C_{zy}}{\gamma_c B_z} \quad (2)$$

where $A_{zx}^{(j)}(C_{zx})$ and $A_{zy}^{(j)}(C_{zy})$ are the perpendicular electron–nuclear (nuclear–nuclear) coupling components. We cancel this term by averaging the double-resonance frequencies measured for the $m_s = \pm 1$ electron-spin projections.

The remaining electron-mediated corrections depend on the angles of the electron–nuclear hyperfine interactions. Because these angles are unknown, we estimate the maximum possible shift for each spin–spin interaction by maximizing over all angles. For our cluster (Fig. 4), most of these maximum possible shifts are small (their average value is ~ 0.03 Hz). In rare cases, the maximum possible correction runs up to 0.6 Hz (see Supplementary Information section IV) but, as the locations of the involved spins are already precisely fixed through strong (>20 Hz) interactions with several other spins, this would have a negligible effect on the obtained structure. Therefore, we can base the structural analysis on dipole–dipole interactions.

3D structure analysis

The 3D structure of the nuclear spins is obtained using the dipole–dipole coupling formula, which relates the zz couplings C_{ij} to the spatial coordinates x, y, z of spins i and j as

$$C_{ij} = \frac{\alpha_{ij}}{\Delta r_{ij}^3} \left[\frac{3(z_j - z_i)^2}{\Delta r_{ij}^2} - 1 \right] \quad (3)$$

where $\Delta r_{ij} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2}$, $\alpha_{ij} = \mu_0 \gamma_i \gamma_j \hbar / (4\pi)$, μ_0 is the permeability of free space, γ_i is the gyromagnetic ratio of nuclear spin i , and \hbar is the reduced Planck constant.

The goal is to minimize the sum of squares $\xi = \sum_{i < j} |\Delta f_{ij}|^2$, where $\Delta f_{ij} = f_{ij} - [C_{ij}/(4\pi)]$ are the residuals and f_{ij} are the measured coupling frequencies. For $M = 27$ spins, there are $3M - 4 = 77$ free coordinates and $M(M - 1)/2 = 351$ pairwise couplings, of which 171 were determined in this work. ξ can in principle be minimized using standard fitting methods; however, tests with randomly generated spin clusters indicate that the initial guess for the coordinates should be within ~ 0.5 Å in order for the fit to converge to the correct solution. For 27 spins, this

corresponds to an intractable $\sim 10^{100}$ possible initial guesses. Instead, we sequentially build the structure by adding spins one by one.

For the diamond-lattice positioning method, we first use the strongest measured coupling to any spin that is already positioned to reduce the position of a new spin to a number of possible lattice coordinates. For each possible coordinate, we then check if the predicted couplings to all other spins satisfy $\Delta f_{ij} < T$, where $T = 1.1$ Hz is a tolerance that is chosen to ensure that all promising configurations are included while maintaining reasonable computation time. Configurations are discarded if they do not satisfy this requirement for one or more of the pairwise couplings. If more than $X_{\text{cutoff}} = 5,000$ possible configurations are identified, only the best X_{cutoff} solutions are kept, according to their ξ values.

For the cubic-lattice positioning method, the same procedure is followed, with the key difference being that the lattice is adaptively generated depending on the strongest coupling to a spin already positioned in the cluster (see Supplementary Information section III). This ensures that in each case the lattice spacing is fine enough to appropriately sample the volume associated with the dipole–dipole coupling between the nuclear spins.

Robustness of the analysis

The method is robust to failure. The problem is generally highly overdetermined, so that discarding the correct configuration because of X_{cutoff} will lead to no solution at all, instead of an erroneous solution. Given enough computational resources, a correct solution is always expected to be found. As a test, we used the cubic-lattice reconstruction method on 17 randomly generated 30-spin clusters with added noise, and no erroneous structures were returned (see Supplementary Information section IV).

Comparison to 1D Ramsey spectroscopy

Extended Data Fig. 2 compares the 1D Ramsey signal with spectra reconstructed using 3D spectroscopy. This comparison illustrates our method's effective resolution improvement and its ability to resolve dense spectra. We note that, apart from the spectral resolution, the signals should not be compared directly, because the Ramsey experiment is difficult to interpret quantitatively. First, the Ramsey signals probably contain contributions from multiple spins, due to both spectral overlap and higher-order contributions^{36–38}. Second, any inadvertent polarization of other spins in the cluster or the environment modifies the spectrum. These effects are difficult to separate from actual nuclear–nuclear couplings, and the fact that the spectra are asymmetric indicates that they play a non-negligible role. Our 3D spectroscopy method resolves these issues.

Finding the position of the NV centre

Because the NV electron wavefunction is not known a priori, we cannot use the electron–nuclear couplings to find the NV position. In particular, DFT calculations³³ indicate that for electron–nuclear couplings in the range observed here, assuming a point–dipole model for the electron spin can lead to large discrepancies and is therefore not justified.

Our approach is to measure the couplings between the ^{13}C spins and the NV nitrogen nuclear spin, for which the point–dipole approximation is accurate. The nitrogen– ^{13}C couplings can be measured using a double-resonance procedure similar to that used for measuring ^{13}C – ^{13}C couplings. We use the nitrogen spin as the probe spin, which gives better spectral resolution owing to its longer coherence time ($T_2 = 2.3(2)$ s)²⁸. We initialize the nitrogen spin in the state $m_i = 0$ using measurement-based initialization³⁴ and manipulate the spin state using RF pulses (m_i , nuclear-spin projection). Extended Data Fig. 4b shows the measured couplings between the nitrogen and ^{13}C spins.

Using the couplings, the nitrogen spin is added to the ^{13}C nuclear-spin cluster using the diamond-lattice positioning method, with $\gamma_j \rightarrow \gamma_n = 2\pi \times 0.3077$ kHz G^{−1} (γ_n is the nitrogen gyromagnetic ratio), in equation (3). Determining the nitrogen lattice site also allows the

vacancy site to be determined on the basis of the known NV distance and the alignment with the magnetic field along z , thereby giving the location and the orientation of the NV centre with respect to the ^{13}C nuclear-spin cluster. The resulting 3D plot of the best solution is shown in Extended Data Fig. 4a. The nitrogen spin coordinate is the same for all 5,000 configurations identified. Extended Data Fig. 4c gives the results of a least-squares fit.

Comparison to DFT

Now that we have independently determined the position of the ^{13}C spins relative to the NV centre, we can compare the hyperfine couplings to DFT calculations without any prior assumptions. In Nizovtsev et al.³³, hyperfine couplings are calculated for 510 lattice sites surrounding the NV centre. Extended Data Fig. 4d shows the lattice positions given in ref.³³, along with the coordinates of the ^{13}C spins found in this work. To match the coordinate frames, the ^{13}C spin coordinates are mirrored ($z \rightarrow -z$) and transformed so that the nitrogen spin is at the origin. Additionally, a scaling factor of 1.02 is applied, which was found by comparing the 510 lattice sites from ref.³³ with the same sites in our work. Five of the 27 spins identified in this work were calculated in ref.³³. The remaining spins cannot yet be compared with DFT calculations. Extended Data Fig. 4e shows the measured electron– ^{13}C hyperfine couplings (see Supplementary Table 1), as well as those predicted in ref.³³, for the five spins. For the DFT results, we take the average of the predicted couplings for the possible C_{3v} symmetric lattice sites. Additionally, we take the negative of the predicted parallel component of the electron–nuclear coupling A_{\parallel} for all spins (a global minus sign is possible because of the unknown orientation of the magnetic field along z).

Data availability

The data that support the findings of this study are available from the corresponding author upon request.

34. Robledo, L. et al. High-fidelity projective read-out of a solid-state spin quantum register. *Nature* **477**, 574–578 (2011).

35. Toyli, D. M., de las Casas, C. F., Christle, D. J., Dobrovitski, V. V. & Awschalom, D. D. Fluorescence thermometry enhanced by the quantum coherence of single spins in diamond. *Proc. Natl Acad. Sci. USA* **110**, 8417–8421 (2013).
36. Taminiau, T. H. et al. Detection and control of individual nuclear spins using a weakly coupled electron spin. *Phys. Rev. Lett.* **109**, 137602 (2012).
37. Kolkowitz, S., Unterreithmeier, Q. P., Bennett, S. D. & Lukin, M. D. Sensing distant nuclear spins with a single electron spin. *Phys. Rev. Lett.* **109**, 137601 (2012).
38. Zhao, N. et al. Sensing single remote nuclear spins. *Nat. Nanotechnol.* **7**, 657–662 (2012).
39. Boss, J. M., Cuijia, K., Zopes, J. & Degen, C. L. Quantum sensing with arbitrary frequency resolution. *Science* **356**, 837–840 (2017).
40. Schmitt, S. et al. Submillihertz magnetic spectroscopy performed with a nanoscale quantum sensor. *Science* **356**, 832–837 (2017).
41. Laraoui, A., Pagliaro, D. & Meriles, C. A. Imaging nuclear spins weakly coupled to a probe paramagnetic center. *Phys. Rev. B* **91**, 205410 (2015).
42. Gullion, T., Baker, D. B. & Conradi, M. S. New, compensated Carr–Purcell sequences. *J. Magn. Reson.* **89**, 479–484 (1990).
43. Warren, W. S. Effects of arbitrary laser or NMR pulse shapes on population inversion and coherence. *J. Chem. Phys.* **81**, 5437–5448 (1984).
44. Zhao, N., Hu, J.-L., Ho, S.-W., Wan, J. T. K. & Liu, R. B. Atomic-scale magnetometry of distant nuclear spin clusters via nitrogen-vacancy spin in diamond. *Nat. Nanotechnol.* **6**, 242–246 (2011).

Acknowledgements We thank V. V. Dobrovitski, T. van der Sar, W. Hahn, M. Scheer and R. Zia for valuable discussions. We thank R. F. L. Vermeulen and R. N. Schouten for assistance with the RF electronics, and M. Eschen for assistance with the experimental setup. We acknowledge support from the Netherlands Organisation for Scientific Research (NWO/OCW) through a Vidi grant, as part of the Frontiers of Nanoscience (NanoFront) programme and as part of the Quantum Software Consortium programme (project number 024.003.037/3368).

Author contributions M.H.A. and T.H.T. devised the experiments. M.H.A. performed the experiments. J.R. developed the 3D structure analysis method. M.H.A., J.R. and T.H.T. analysed the data. M.H.A., J.R., M.J.D. and C.E.B. prepared the experimental apparatus. C.E.B. and J.R. developed the RF electronics. H.P.B. and M.A.B. performed preliminary experiments. M.H.A., M.J.D., J.R., C.E.B. and T.H.T. developed the magnetic-field alignment procedure and the ^{14}N echo spectroscopy. M.M. and D.J.T. grew the diamond sample. M.H.A., J.R. and T.H.T. wrote the manuscript with input from all authors. T.H.T. supervised the project.

Competing interests The authors declare no competing interests.

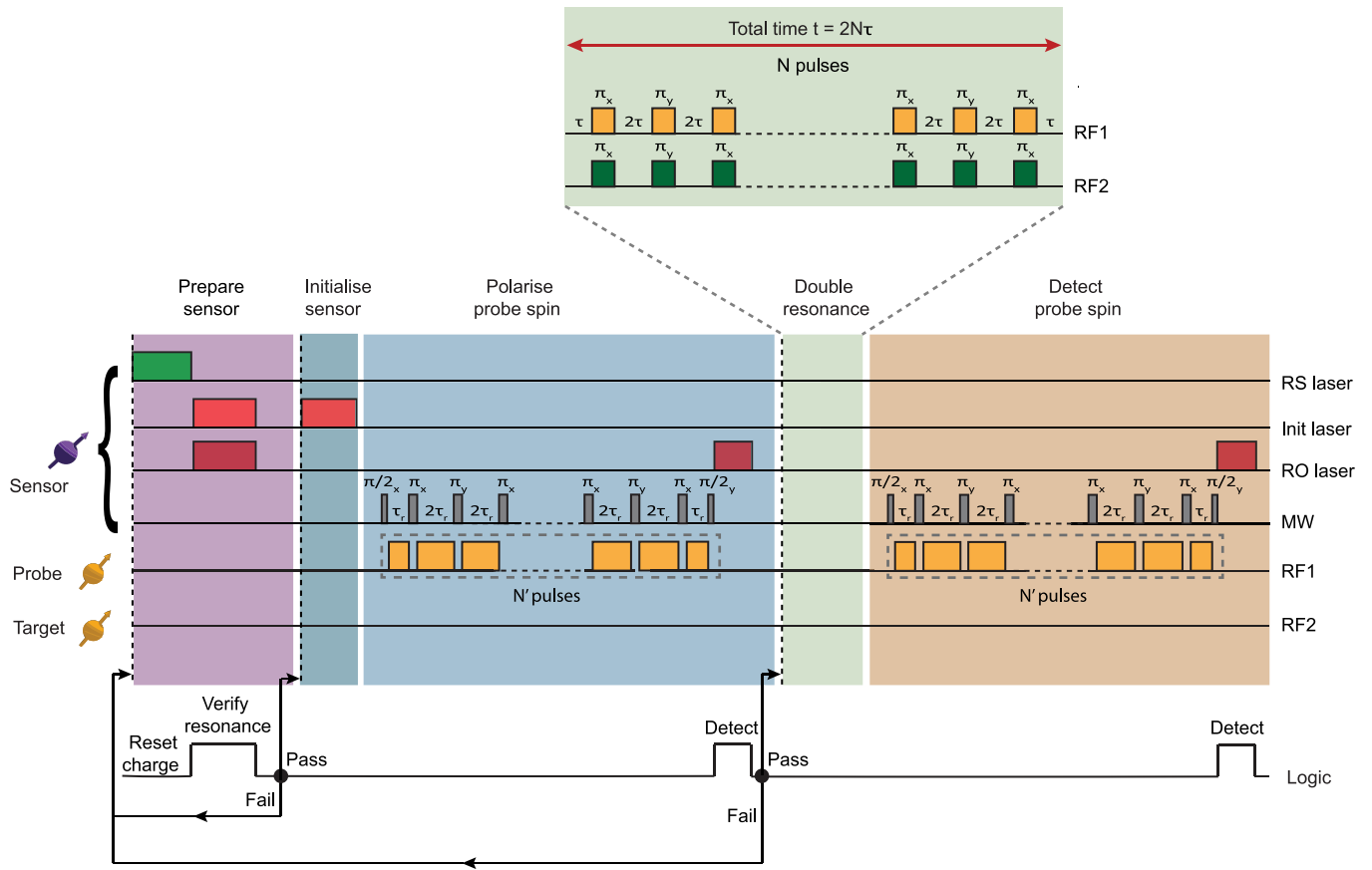
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1834-7>.

Correspondence and requests for materials should be addressed to T.H.T.

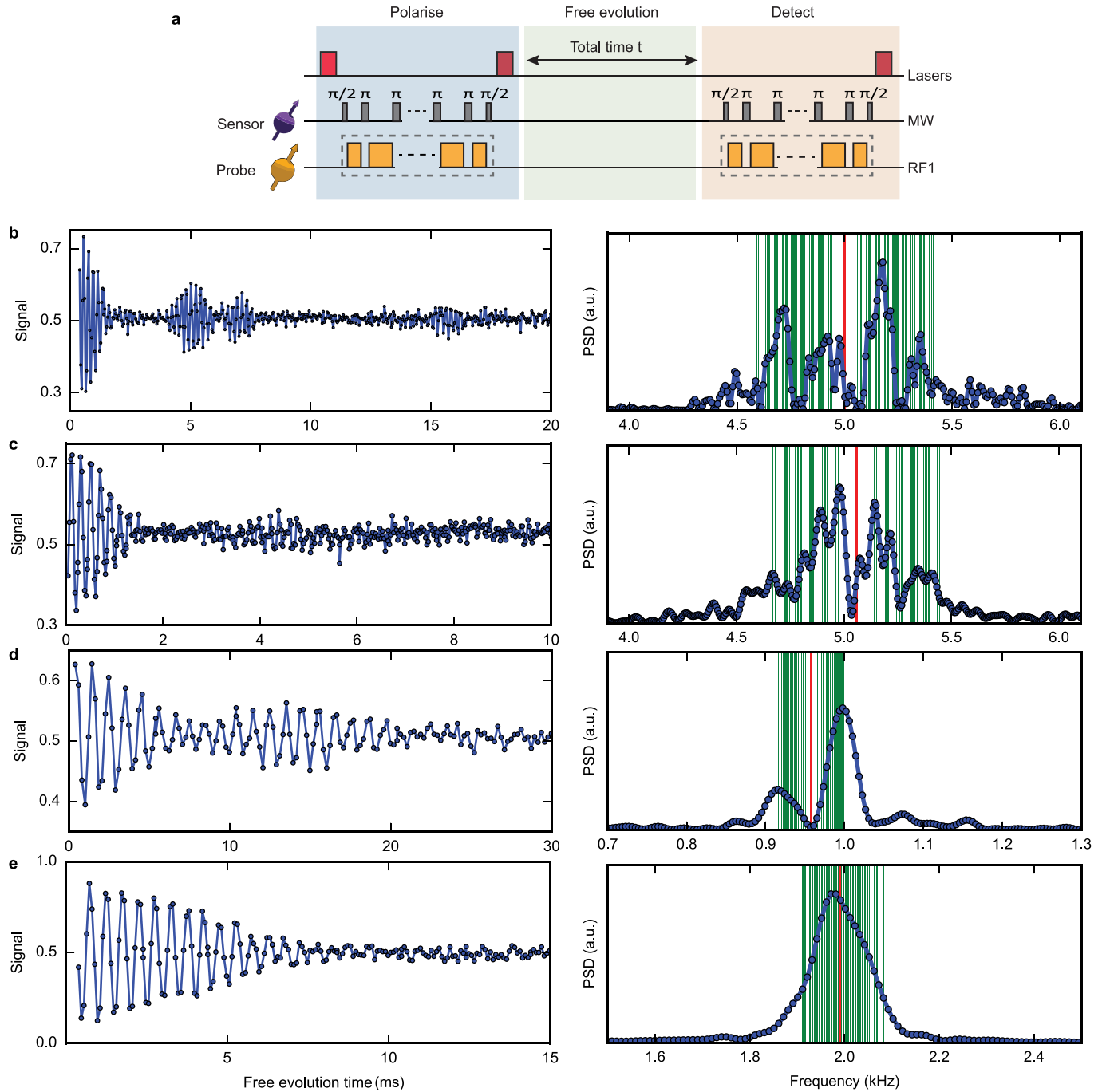
Peer review information *Nature* thanks Nir Bar-Gill, Vidya Praveen Bhallamudi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | Experimental sequence. The pulse sequence consists of five parts: sensor preparation, sensor initialization, polarization of the probe spin, double resonance, and detection of the probe spin. Sensor preparation: the NV centre is prepared by excitation with two 637-nm lasers for 150 μ s and counting of the detected photons (read-out (RO) laser resonant with the E_x transition and initialization (Init) laser resonant with the E' transition)^{24,34}. If the number of photons exceeds a certain threshold, the NV is in the negative charge state and resonant with both lasers, and we proceed to the next step. If not, we apply a 515-nm laser (charge reset (RS) laser, 1 ms) and repeat the process^{24,34}. Sensor initialization: the NV electron spin is initialized in the $m_s = 0$ state through spin pumping (Init laser, 100 μ s)³⁴. Polarization of the probe spin: first, the NV sensor is brought into a superposition state using a $\pi/2$ pulse. Then, a dynamical decoupling sequence of N' equally spaced π pulses of the

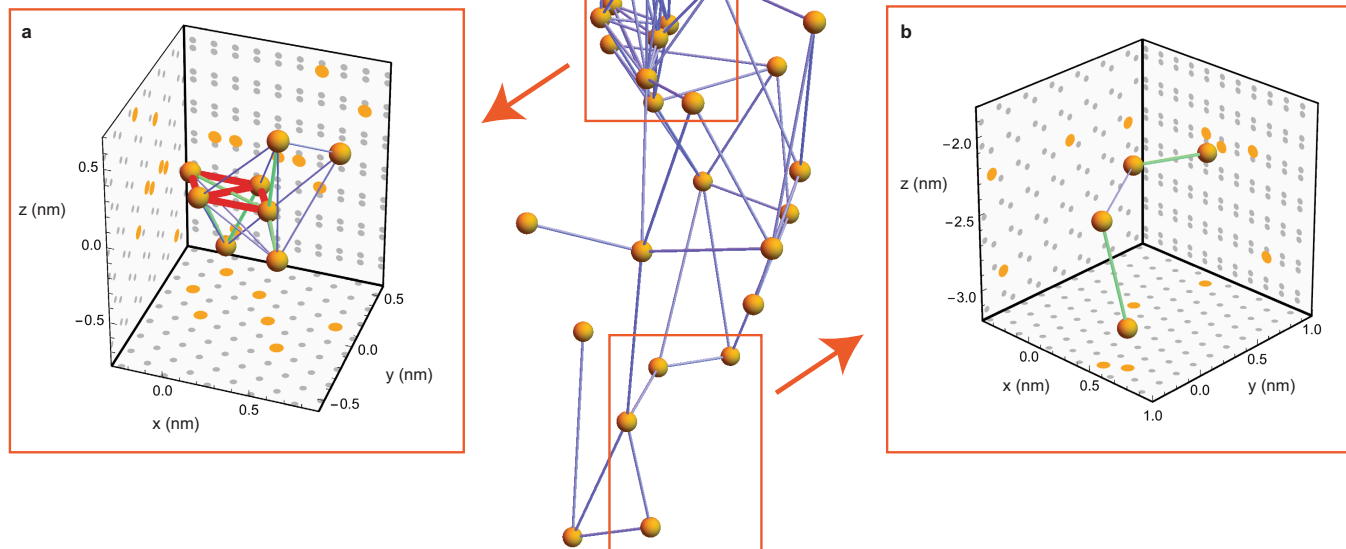
form $(\tau_r - \pi - \tau_r)^{N'}$ is applied to the electron spin. This sequence correlates the state of the nuclear spin(s) with the phase of the electron spin. We use two different sequences (see Methods). For sequence B, the microwave (MW) π pulses are interleaved with RF pulses (RF1) that resonantly drive the probe spin(s) (dashed box); see Bradley et al.²⁸ for details. A second $\pi/2$ pulse maps the electron phase to the population and the electron spin is read out (RO laser). Double resonance: N echo pulses are applied simultaneously on the probe spin(s) (RF1) and the target spin(s) (RF2), so that the coupling between these spins is isolated. To mitigate pulse errors, we alternate the phases of the pulses following the XY8 scheme⁴². Detection of the probe spin: the detection sequence is the same as the polarization sequence except for the final RO laser pulse, which is applied for 10 μ s and with higher power.



Extended Data Fig. 2 | Ramsey experiments and reconstructed spectra.

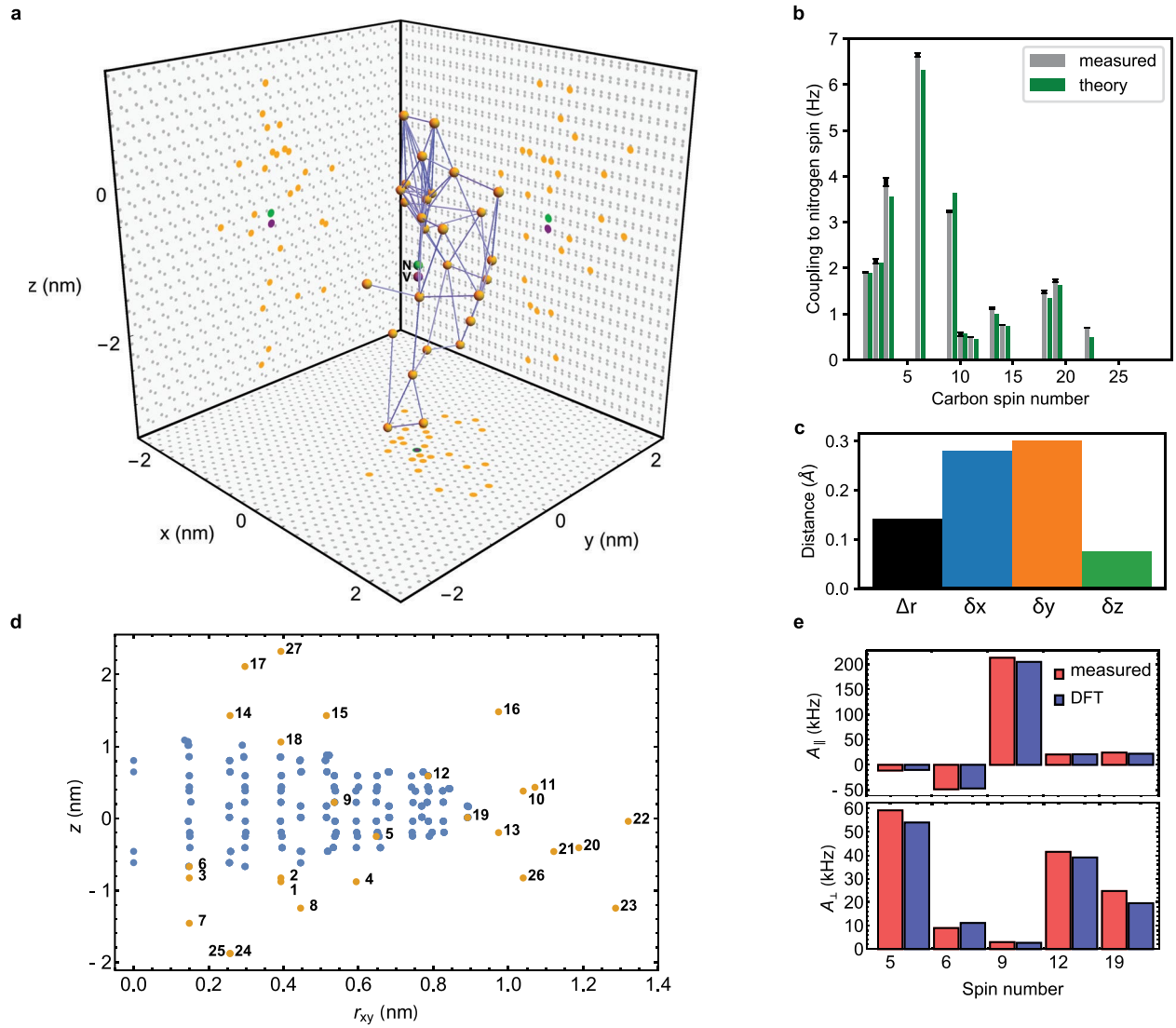
a, Schematic of the pulse sequence used in the Ramsey experiment (equivalent to correlation spectroscopy); see Methods and Extended Data Fig. 1 for details. **b**, Ramsey signal for C2 and the corresponding PSD (5 kHz detuning). The red line represents the central frequency f_0 . Green lines are the 2^7 frequencies

based on the seven strongest coupling strengths extracted from our high-resolution double-resonance spectroscopy (Supplementary Table 4). These frequencies are given by $f_0 \pm f_1 \pm f_2 \pm f_3 \pm f_4 \pm f_5 \pm f_6 \pm f_7$, where f_1 to f_7 are the seven largest measured coupling strengths for C2. **c–e**, The same experiment for C3 (c; -5 kHz detuning), C15 (d; -1 kHz detuning) and C5 (e; -2 kHz detuning).



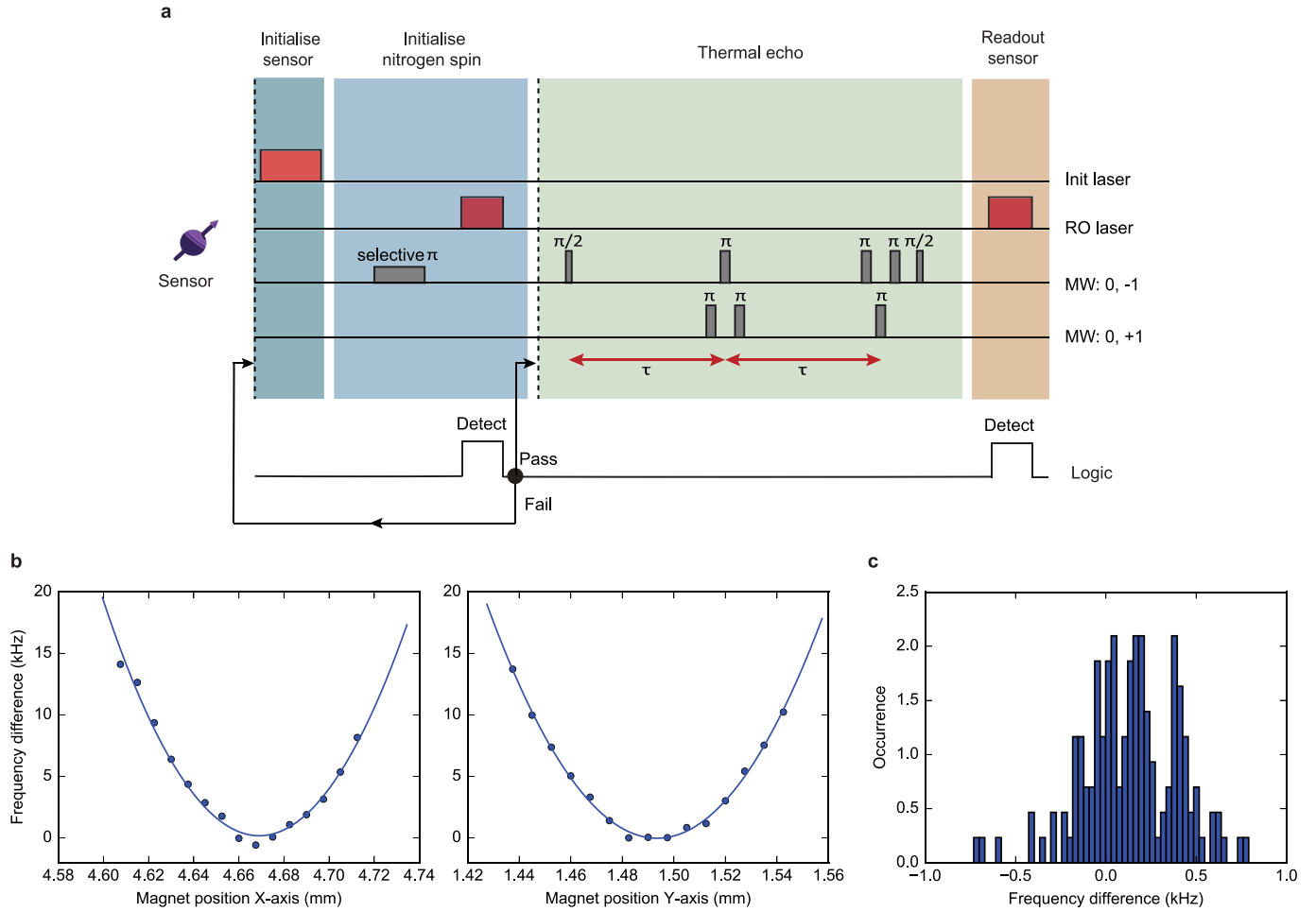
Extended Data Fig. 3 | Strongly coupled subclusters. 3D plots showing the structure of two strongly coupled subclusters (orange boxes) within the larger cluster (shown in the centre). Ramsey measurements performed on spins within these subclusters show clear beating signals within their dephasing time

T_2^* (see, for example, Extended Data Fig. 2). **a**, Eight-spin subcluster. **b**, Four-spin subcluster. Couplings above 3 Hz are marked blue, above 20 Hz green and above 50 Hz red. Grey points show the 2D projections of the diamond-lattice coordinates.



Extended Data Fig. 4 | Finding the position of the NV centre. **a**, 3D plot showing the 27-nuclear-spin cluster shown in Fig. 4, with the position of the nitrogen spin (green) and vacancy (purple) lattice sites calculated from the measured nitrogen- ^{13}C couplings. The grey dots show the 2D projections of the diamond-lattice coordinates. **b**, Bar plot showing the measured couplings f_{iN} between ^{13}C spin i and the nitrogen spin (grey), as well as the theoretically calculated couplings $|C_{iN}|/4\pi$ (green). Error bars are one standard deviation. See Supplementary Table 4 for the numerical values. **c**, Bar plots of Δr for the fitted position of the nitrogen spin (black) and fit errors δx (blue), δy (orange)

and δz (green), where the ^{13}C spins are fixed at the diamond-lattice solution. **d**, Plot of z versus $r_{xy} = \sqrt{x^2 + y^2}$ for all lattice positions used in the DFT calculation from Nizovtsev et al. ³³ (blue) and for the appropriately transformed ^{13}C coordinates found in this work (orange). Spins 5, 6, 9, 12 and 19 match a DFT lattice position, whereas the rest of the identified spins are outside the 510 lattice sites simulated. **e**, Measured electron- ^{13}C parallel (top) and perpendicular (bottom) hyperfine couplings for the five spins that are within the DFT calculation volume (red; from Supplementary Table 1), compared with the DFT results from ref. ³³ (blue).



Extended Data Fig. 5 | Aligning the magnetic field using a thermal echo sequence. **a**, Pulse sequence used for the thermal echo measurement³⁵. The electron spin is prepared in a superposition of the states $m_s = 0$ and $m_s = -1$ in the first half of the sequence and then swapped to a superposition of $m_s = 0$ and $m_s = +1$ for the second half, using a sequence of three closely spaced π pulses. By sweeping τ , the average frequency $f_{TE} = (f_{+1} + f_{-1})/2$ is obtained, which is minimized when $B_{\perp} = 0$. $f_{\pm 1}$ are the $m_s = 0 \leftrightarrow m_s = \pm 1$ transition frequencies. The NV nitrogen spin is initialized in $m_i = 0$ (ref. ³⁴). **b**, Magnetic-field alignment by scanning the magnet position in two orthogonal directions. The obtained

thermal echo frequencies are fitted with a parabolic function to find the optimum position (that is, minimal f_{TE}). The plots show the frequency difference $f_{TE} - 2.877652$ GHz. **c**, The magnet is placed at the optimum position and the measurement is repeated 200 times (over a 10-h period). The obtained average frequency difference is 0.13 kHz, with a standard deviation of 0.27 kHz, which is consistent with the statistical measurement error. Therefore, the total uncertainty for the magnet alignment is -0.4 kHz, which corresponds to a perpendicular field of 0.5 G or a misalignment angle of 0.07° .

Prediction and observation of an antiferromagnetic topological insulator

<https://doi.org/10.1038/s41586-019-1840-9>

Received: 20 September 2018

Accepted: 18 September 2019

Published online: 18 December 2019

M. M. Otrokov^{1,2,3,4*}, I. I. Klimovskikh⁴, H. Bentmann⁵, D. Estyunin⁴, A. Zeugner⁶, Z. S. Aliev^{7,8}, S. Gaß⁹, A. U. B. Wolter⁹, A. V. Koroleva⁴, A. M. Shikin⁴, M. Blanco-Rey^{3,10}, M. Hoffmann¹¹, I. P. Rusinov^{4,12}, A. Yu. Vyazovskaya^{4,12}, S. V. Ereemeev^{4,12,13}, Yu. M. Koroteev^{12,13}, V. M. Kuznetsov¹², F. Freyse¹⁴, J. Sánchez-Barriga¹⁴, I. R. Amiraslanov⁷, M. B. Babanly¹⁵, N. T. Mamedov⁷, N. A. Abdullayev⁷, V. N. Zverev¹⁶, A. Alfonsov⁹, V. Kataev⁹, B. Büchner^{9,17}, E. F. Schwier¹⁸, S. Kumar¹⁸, A. Kimura¹⁹, L. Petaccia²⁰, G. Di Santo²⁰, R. C. Vidal⁵, S. Schatz⁵, K. Kißner⁵, M. Ünzelmann⁵, C. H. Min⁵, Simon Moser²¹, T. R. F. Peixoto⁵, F. Reinert⁵, A. Ernst^{11,22}, P. M. Echenique^{1,3,10}, A. Isaeva^{9,17} & E. V. Chulkov^{1,3,4,10*}

Magnetic topological insulators are narrow-gap semiconductor materials that combine non-trivial band topology and magnetic order¹. Unlike their nonmagnetic counterparts, magnetic topological insulators may have some of the surfaces gapped, which enables a number of exotic phenomena that have potential applications in spintronics¹, such as the quantum anomalous Hall effect² and chiral Majorana fermions³. So far, magnetic topological insulators have only been created by means of doping nonmagnetic topological insulators with 3d transition-metal elements; however, such an approach leads to strongly inhomogeneous magnetic⁴ and electronic⁵ properties of these materials, restricting the observation of important effects to very low temperatures^{2,3}. An intrinsic magnetic topological insulator—a stoichiometric well ordered magnetic compound—could be an ideal solution to these problems, but no such material has been observed so far. Here we predict by ab initio calculations and further confirm using various experimental techniques the realization of an antiferromagnetic topological insulator in the layered van der Waals compound MnBi₂Te₄. The antiferromagnetic ordering that MnBi₂Te₄ shows makes it invariant with respect to the combination of the time-reversal and primitive-lattice translation symmetries, giving rise to a \mathbb{Z}_2 topological classification; $\mathbb{Z}_2 = 1$ for MnBi₂Te₄, confirming its topologically nontrivial nature. Our experiments indicate that the symmetry-breaking (0001) surface of MnBi₂Te₄ exhibits a large bandgap in the topological surface state. We expect this property to eventually enable the observation of a number of fundamental phenomena, among them quantized magnetoelectric coupling^{6–8} and axion electrodynamics^{9,10}. Other exotic phenomena could become accessible at much higher temperatures than those reached so far, such as the quantum anomalous Hall effect² and chiral Majorana fermions³.

The first reference to MnBi₂Te₄ as a stable chemical compound was in 2013 when it was synthesized in the powder form¹¹. The trigonal structure (space group $R\bar{3}m$) of MnBi₂Te₄ comprises septuple-layer blocks stacked along the [0001] direction and bound to each other by van der Waals forces (Fig. 1a). High quality MnBi₂Te₄ single crystals were grown at TU Dresden (hereafter, D samples; Fig. 2a) and at ASOIU Baku (B samples; Fig. 2b). Our single-crystal X-ray diffraction experiments confirm the same lattice symmetry reported previously¹¹. Because the magnetism of MnBi₂Te₄ has not been experimentally investigated, we began our study by calculating the exchange-coupling parameters/ J from first principles (Fig. 1b, c). Among the intralayer interactions/ J^{\parallel} the one between the nearest neighbours in the Mn layer is clearly dominant ($J^{\parallel}_{0,1} \approx 0.09$ meV per μ_B^2 where μ_B is the Bohr magneton), and the inter-

actions with more distant neighbours are an order of magnitude weaker, whereupon a ferromagnetic ordering is expected within each septuple-layer block of MnBi₂Te₄. In contrast, the interlayer coupling constants J^{\perp} are mostly negative, resulting in a negative $J^{\perp}_0 \approx \sum_{j \neq 0} J^{\perp}_{0,j}$ coefficient equal to -0.022 meV per μ_B^2 , which means that the overall coupling between neighbouring Mn layers is antiferromagnetic (AFM) (Fig. 1d). The magnetic anisotropy energy is positive and equal to 0.225 meV per Mn atom, indicating the easy axis with an out-of-plane orientation of the local magnetic moments of $\pm 4.607\mu_B$. Monte Carlo simulations confirm the interlayer AFM structure with a Néel temperature (T_N) of three-dimensional ordering of 25.4 K (Extended Data Fig. 1).

Given the magnetic ground state, the bulk electronic structure was calculated. As shown in Fig. 1f, the system is insulating, the fundamental

The list of affiliations appears at the end of the paper.

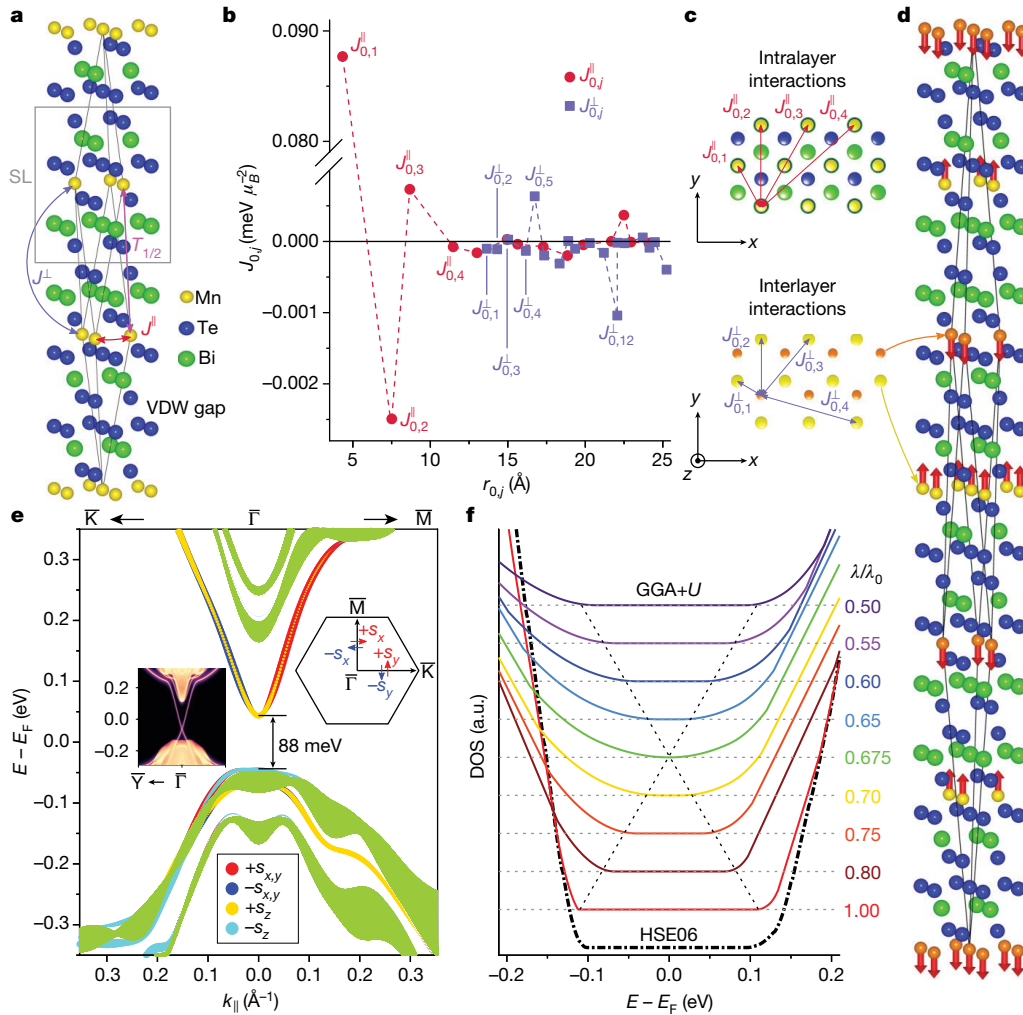


Fig. 1 | Theoretical insights into the crystal, magnetic and electronic structure of MnBi_2Te_4 . **a**, Crystal structure of the trigonal MnBi_2Te_4 with yellow, blue and green spheres showing Mn, Te and Bi atoms, respectively. The ‘nonmagnetic’ rhombohedral primitive unit cell is indicated by grey lines and the $T_{1/2}$ translation is shown in magenta. VdW, van der Waals; SL, septuple layer. **b**, Calculated exchange constants J_{0j} [meV per μ_B^2] for the intralayer (J^{\parallel} , red circles) and interlayer (J^{\perp} , light blue squares) pair interactions as a function of the Mn–Mn distance, r_{0j} [Å]. **c**, Schematic top view representation of magnetic interactions in the Mn layer (top) and between two neighbouring Mn layers (bottom). **d**, Magnetic unit cell corresponding to the interlayer AFM state. The arrows on atoms denote the local magnetic moments. **e**, Spin-resolved electronic structure of the MnBi_2Te_4 (0001) surface. The size of the coloured circles that comprise the data reflects the value and sign of the Cartesian projections of the spin-vector. **f**, Red and blue circles correspond to the positive and negative s_x and s_y components (perpendicular to k_{\parallel}), and yellow

and cyan to the out-of-plane components $+s_z$ and $-s_z$. The green areas correspond to the bulk bandstructure projected onto the surface Brillouin zone. Left inset, The tight-binding calculated electronic bandstructure of the S -preserving (10 $\bar{1}$ 1) surface (see also Extended Data Fig. 2). The regions with a continuous spectrum correspond to the three-dimensional bulk states projected onto the two-dimensional Brillouin zone. Right inset, the (0001) surface Brillouin zone showing the high symmetry directions along which the bandstructure shown in the main panel was calculated. Colour arrows show the Cartesian projections of the spin vector. **f**, Total DOS of bulk MnBi_2Te_4 calculated for the interlayer AFM state shown in **d** using the HSE06 exchange–correlation functional (dash-dot black line) and the GGA+ U approach (solid lines). For GGA+ U , the evolution of the DOS with the change of the SOC constant λ is shown (coloured lines). The thin dashed lines mark a zero DOS level for each dataset, and their intersections with the inclined lines approximately mark the bulk bandgap edges for each λ/λ_0 value. a.u., arbitrary units.

bandgap value, determined from the GGA+ U calculation (the generalised gradient approximation with a Hubbard U correction for the Mn 3d states; see Methods section) of the density of states (DOS), is around 220 meV. To determine whether the gap is negative (inverted), we performed the DOS calculations decreasing the spin–orbit coupling (SOC) constant λ stepwise from its natural value λ_0 to $\lambda = 0.5\lambda_0$. It was found that at $\lambda/\lambda_0 \approx 0.675$ the gap is closed and at other values of λ/λ_0 it is non-zero, which points towards a nontrivial topology of MnBi_2Te_4 .

The \mathbb{Z}_2 classification of AFM insulators was introduced⁶ on the basis of the $S = \Theta T_{1/2}$ symmetry, which is inherent to MnBi_2Te_4 (Θ is the time-reversal symmetry and $T_{1/2}$ is the primitive lattice translation symmetry). We find $\mathbb{Z}_2 = 1$, which classifies MnBi_2Te_4 as an AFM topological insulator.

The implication of the bulk bandgap inversion in a topological insulator is seen at its surface, where the topological phase manifests itself by the appearance of the topological surface state (TSS). In the case of nonmagnetic topological insulators, this surface state is gapless¹²; however, at the (0001) surface of the AFM topological insulator MnBi_2Te_4 we find a 88-meV-wide bandgap (Fig. 1e). Here, the S symmetry is broken and the out-of-plane magnetization of the near-surface ferromagnetic layer opens the Dirac point gap⁶. In contrast, the S -preserving surface is gapless, as expected for an AFM topological insulator (inset to Fig. 1e and Extended Data Fig. 2).

Temperature- and field-dependent magnetization measurements performed on D samples (Fig. 2c, e) establish a three-dimensional AFM order below $T_N = 24.2(5)$ K, in agreement with the prediction by

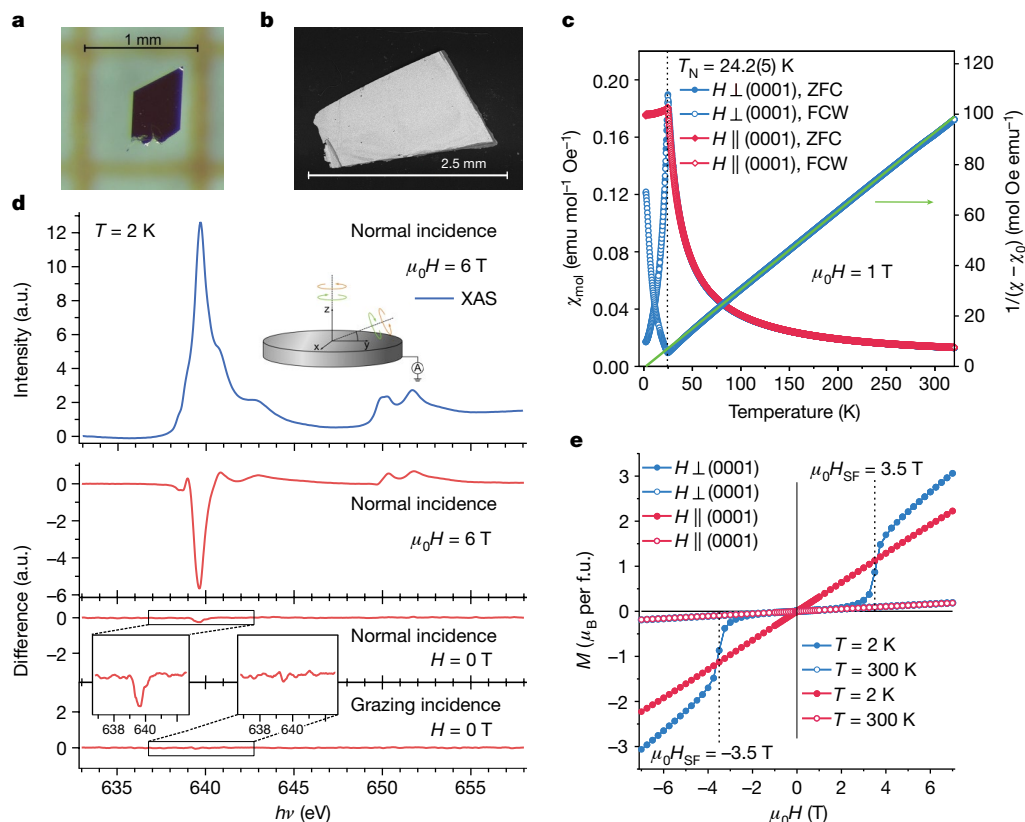


Fig. 2 | MnBi₂Te₄ single crystals and their magnetic properties. **a, b**, MnBi₂Te₄ single crystals: D sample (**a**, optical microscope image) and B sample (**b**, scanning electron microscope image). **c**, Magnetic susceptibility (left axis) of phase-pure MnBi₂Te₄ as a function of temperature measured in an external magnetic field of $\mu_0 H = 1$ T in zero-field-cooled (ZFC) and field-cooled-warming (FCW) conditions, alongside the temperature-dependent reciprocal Curie–Weiss fit to the high-temperature data ($\chi_0 = 0.0028(3)$ e.m.u. mol⁻¹ Oe⁻¹, 1 oersted (Oe) = $1,000/4\pi$ A m⁻¹, 1 electromagnetic unit (e.m.u.) = $4\pi \times 10^{-6}$ m³ mol⁻¹; details in text). **d**, XMCD measurements for MnBi₂Te₄ at the Mn L_{2,3} edge. Inset, A sketch of the experiment. The external magnetic field is

applied along the direction of light incidence. Top, The sum (X-ray absorption spectroscopy signal) and middle, the difference (XMCD signal) between the right (R) and left (L) X-ray absorption intensities I_R and I_L measured with right and left circularly polarized light in normal incidence at $\mu_0 H = 6$ T. Bottom, The XMCD signal in remanence—that is, at $H = 0$ T—for normal and grazing light incidence, measured after switching off an external field of $\mu_0 H = 6$ T along the respective directions. Insets show a magnification of the L₃ dichroism in remanence. **e**, Field-dependent magnetization curves for the two directions, measured at 2 K (blue) and 300 K (red). f.u., formula unit; H_{SF} , spin-flop magnetic field.

Monte Carlo simulations and results of the resistivity measurements (Extended Data Fig. 3a). Below T_N , a strongly anisotropic magnetic susceptibility χ is observed, which decreases more steeply for the magnetic field $H \perp (0001)$. No splitting between zero-field-cooled and field-cooled-warming curves was found. The paramagnetic regime above T_N was fitted with a modified Curie–Weiss law, $\chi(T) = \chi_0 + C/(T - \theta_{CW})$, in the 100 K to 250 K range. Here, χ_0 is the temperature-independent magnetic susceptibility of both diamagnetic closed electron shells and a Pauli paramagnetic contribution resulting from some degree of metallicity in this material (see below). $C/(T - \theta_{CW})$ accounts for a temperature-dependent Curie–Weiss susceptibility of the Mn local magnetic moments. The fitted effective paramagnetic moment of $5.0(2)\mu_B$ is in rough agreement with the high-spin configuration of Mn²⁺ ($S = 5/2$), and a small and positive value of the Curie–Weiss temperature ($\theta_{CW} = 3(3)$ K) strongly depends on the fitted χ_0 contribution. The $M(H)$ curve acquired below T_N for $H \perp (0001)$ shows an indicative spin-flop transition at $\mu_0 H_{SF} \approx 3.5$ T (Fig. 2e), which is in line with an out-of-plane easy axis of the staggered magnetization.

We further performed X-ray magnetic circular dichroism (XMCD) experiments at the Mn L_{2,3} absorption edge (D samples; Fig. 2d). The data were acquired in total electron yield mode with a probing depth of typically only a few nanometres. The XMCD signal obtained at an external field of 6 T and in normal light incidence verifies the magnetic

polarization of the Mn ions. After removing the external field ($H = 0$) the signal collapses, as expected for an AFM ordering. However, a small residual signal is observed in remanence, indicating a finite net out-of-plane polarization within the probed volume of the sample. This residual signal appears to be inconsistent with an AFM intralayer coupling where the orientation of the moments within a Mn layer alternates on the atomic scale. However, for ferromagnetic intralayer coupling, the first septuple layer, which is preferentially probed in the total electron yield mode, is expected to be composed of mesoscopic domains with the magnetization pointing in or out of the surface plane. We attribute the residual XMCD signal in remanence to a preferential sampling of one domain type with the micrometre-sized synchrotron beam spot. This supports our first-principles calculations, which predict a ferromagnetic ordering within individual septuple layers. Performing the same experiment in grazing light incidence, that is, with sensitivity to in-plane magnetization, we observe no finite polarization in remanence.

The scope of experimental evidence (Fig. 2) enables us to identify MnBi₂Te₄ as an interlayer antiferromagnet, in which ferromagnetic Mn layers are coupled antiparallel to each other and the easy axis of staggered magnetization is perpendicular to the layers (Fig. 1d).

We studied the surface electronic structure of MnBi₂Te₄ (0001) using angle-resolved photoemission spectroscopy (ARPES). The intensity

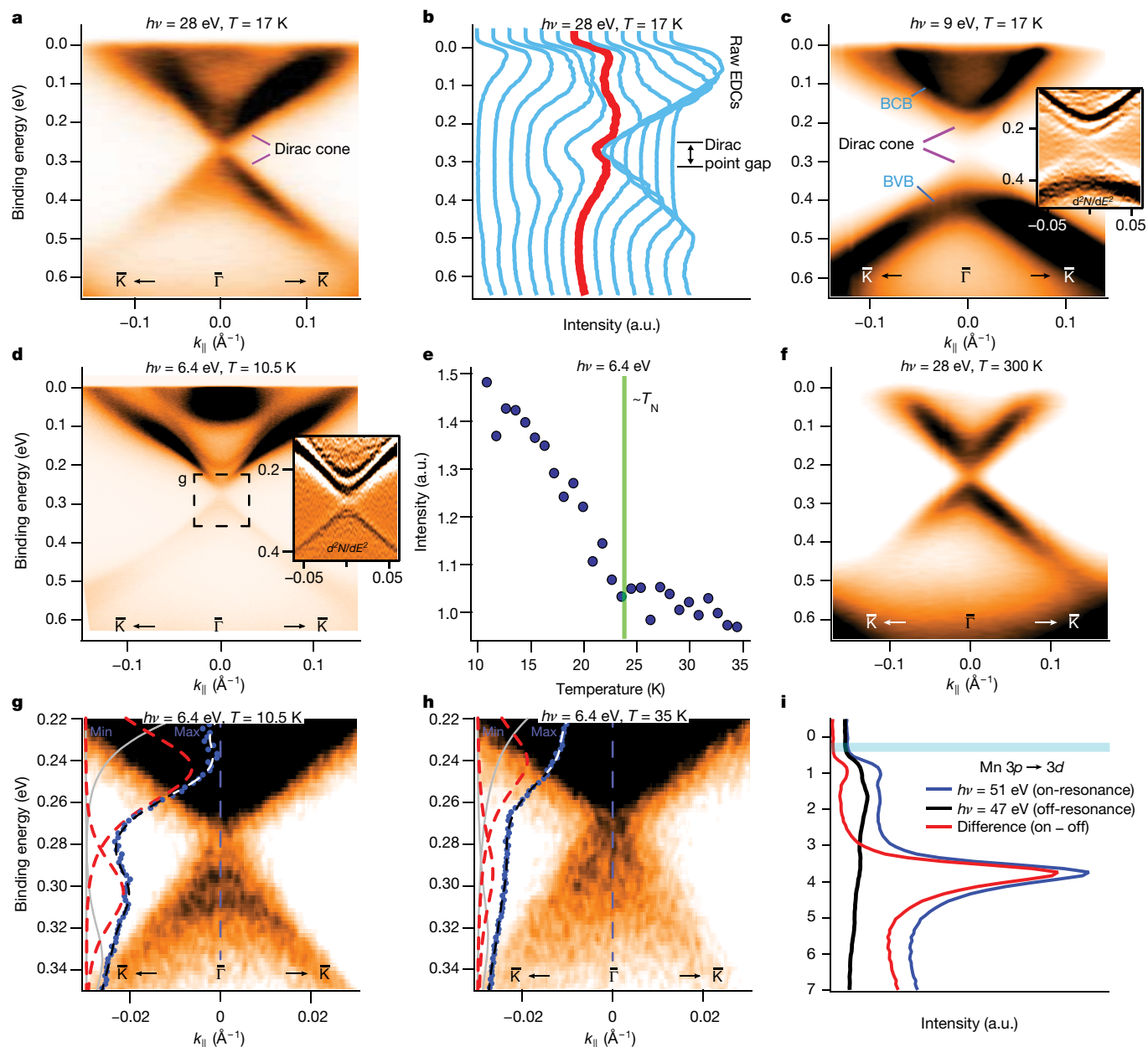


Fig. 3 | Photoemission spectroscopy insight into the surface and bulk bandstructure of MnBi_2Te_4 . **a**, Dispersion of $\text{MnBi}_2\text{Te}_4(0001)$ measured at 17 K with a photon energy of 28 eV. **b**, EDC representation of the data shown in **a**. The red curve marks the EDC at the $\bar{\Gamma}$ -point. **c**, Dispersion of $\text{MnBi}_2\text{Te}_4(0001)$ measured at the same temperature with a photon energy of 9 eV (which is more bulk sensitive). Inset, the corresponding second derivative ($d^2N(E)/dE^2$). BCB, bulk conduction band; BVB, bulk valence band. **d**, The same as **c**, but measured with laser photon energy of 6.4 eV, $T = 10.5$ K, and a different sample. Inset, The corresponding second derivative. The dashed rectangle around the Dirac point marks the region that is magnified in **g**. **e**, Temperature dependence of the Dirac surface state photoemission intensity calculated as a sum of the intensities of the lower and upper parts of the cone at the $\bar{\Gamma}$ -point (see Methods for details). **f**, ARPES image acquired at 300 K ($h\nu = 28$ eV). **g**, **h**, Magnifications

map measured near the Brillouin zone centre at a temperature of 17 K is shown in Fig. 3a ($h\nu = 28$ eV, where h is the Planck constant and ν is the photon frequency; B sample). Two almost linearly dispersing bands form a Dirac-cone-like structure with strongly reduced intensity at the crossing point. The energy distribution curves (EDCs) reveal an energy gap of about 70 meV at the $\bar{\Gamma}$ -point that separates the upper and lower parts of the cone (Fig. 3b). A similar result was obtained for the D

of the Dirac point gap region of the ARPES maps taken at 10.5 K (**g**) and 35 K (**h**), with the fitted EDC spectra at the $\bar{\Gamma}$ -point overlaid. The raw data, resulting fitted curves, and their decompositions with Voigt peaks are shown by blue circles, black and white lines, and grey and red lines, respectively. Red (grey) lines indicate the peaks attributed to the gapped Dirac cone state (bulk bands). The results shown in **a–h** were acquired on B samples. **i**, Resonant valence-band spectra of MnBi_2Te_4 taken at the Mn 3p–3d absorption edge (D samples). On- and off-resonance spectra were obtained at $h\nu = 51$ eV and $h\nu = 47$ eV, respectively. The difference between these spectra approximately reflects the density of the Mn 3d states, showing a main peak near 3.8 eV and an additional feature near 1 eV. The energy range of the bulk energy gap is marked by the horizontal cyan region.

samples (Extended Data Fig. 4). These results agree with those of the $\text{MnBi}_2\text{Te}_4(0001)$ surface bandstructure calculations (see Fig. 1e).

Next, we performed extensive ARPES measurements with different photon energies. At $h\nu = 9$ eV (Fig. 3c), the features that are not seen at $h\nu = 28$ eV show a pronounced spectral weight: namely, the intense electron- and hole-like bands coming to the $\bar{\Gamma}$ -point at binding energies of about 0.17 eV and 0.4 eV, respectively. A comparison with the

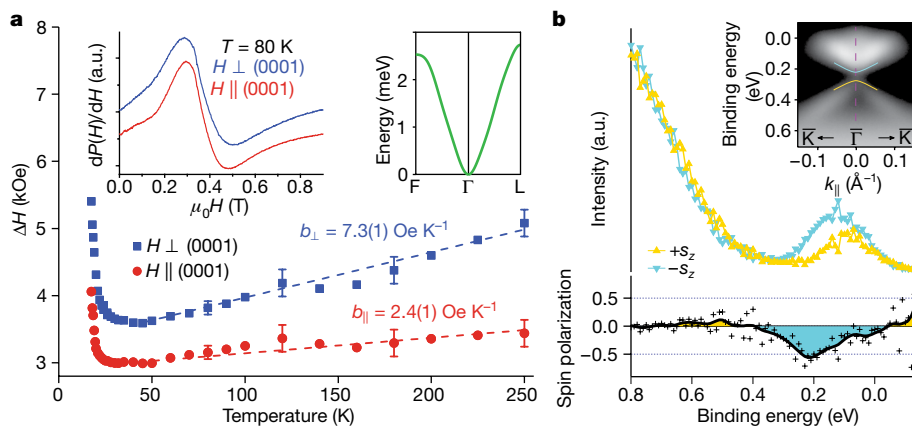


Fig. 4 | Spin characterization of bulk and surface electronic structure of MnBi_2Te_4 . **a**, Temperature dependence of the width ΔH of the ESR signal for two directions of the magnetic field. Dashed lines are linear fits to the $\Delta H(T)$ dependence, yielding the Korringa slopes $b_{\perp,||} = d[\Delta H(T)]/dT$. The error bars are derived from the fitting procedure of ESR line shapes. The larger error bars above approximately 100 K are due to the occurrence of a parasitic signal in the ESR spectrum overlapping with the main line. Left inset, Typical ESR spectra (the field derivative of the microwave absorption $dP(H)/dH$) of Mn^{2+} in MnBi_2Te_4 at $\nu = 9.6$ GHz. The measurements were done on D samples. See Methods for more information about ESR measurements of MnBi_2Te_4 .

Right inset, Magnon spectrum calculated ab initio for bulk MnBi_2Te_4 in the paramagnetic state (paramagnons). **b**, Room-temperature spin-resolved ARPES spectra measured at the Γ -point with respect to the out-of-plane spin quantization axis. The out-of-plane spin polarization is presented below the corresponding spin-up and spin-down spectra (cross symbols). Inset, Spin-integrated ARPES spectrum taken along the $\text{K}-\Gamma-\text{K}$ direction. Yellow and cyan curves show the location of the gapped Dirac cone. The data were obtained for B samples at a photon energy of 6 eV. See Extended Data Fig. 10 for more spin-resolved ARPES data.

theoretically calculated bulk-projected bandstructure enables us to identify these bands as the bulk conduction and valence bands, respectively. The analysis of the $\bar{\Gamma}$ -point EDC shows that both valence and conduction bands can be fitted with two peaks (Extended Data Fig. 5), in agreement with the result of our calculations, showing two bulk bands with a weak k_z dispersion both below and above the Fermi level (Fig. 1e). On the basis of the photoemission measurements, we estimate the bulk bandgap to be close to 200 meV, again in agreement with the calculated values. The second-derivative data in the inset to Fig. 3c provide further insight: along with the bulk bands the gapped Dirac cone is also seen. We note that the Fermi level cuts the conduction band, indicating the n-doped character of the samples, which is consistent with the results of the Hall measurements (see Extended Data Fig. 3b). If we artificially drive MnBi_2Te_4 into a topologically trivial phase by decreasing the SOC strength in a calculation, a strong discrepancy arises between the theory and experiment. Namely, in the trivial phase there are no surface states in the fundamental bulk bandgap (Extended Data Fig. 6). By contrast, the ARPES data acquired at different photon energies (Extended Data Fig. 7) unambiguously confirm the gapped Dirac cone to be a surface state, in agreement with the calculated (0001) surface bandstructure of the MnBi_2Te_4 AFM topological insulator (Fig. 1e).

Thus, by virtue of the good agreement between the theory and experiment, we conclude that the fundamental bandgap of MnBi_2Te_4 is inverted. Our data also confirm the trigonal structure (space group $R\bar{3}m$) and interlayer AFM order, whereby the S -symmetry is also confirmed. These results confirm the $Z_2 = 1$ three-dimensional AFM topological insulator state of MnBi_2Te_4 below its Néel temperature.

The ARPES measurements above T_N were performed next. We note that increasing the temperature does not lead to the Dirac point gap closing at MnBi_2Te_4 (0001); see Fig. 3f. Nevertheless, we observe a temperature dependence of a linear dichroism in the Dirac cone intensity that indicates that the AFM order influences the TSS (see Extended Data Fig. 8). Moreover, directly below T_N , the intensity of the Dirac cone starts to grow abruptly when the temperature decreases (Fig. 3e and Extended Data Fig. 9; laser ARPES with $h\nu = 6.4$ eV), signalling the paramagnet–antiferromagnet transition and showing a clear response of the surface electronic structure to the AFM ordering. In Fig. 3d, the

photoemission intensity map acquired at $T = 10.5$ K is shown. A closer look at these data around the Dirac point gap (Fig. 3g) and their comparison to the analogous spectrum taken at 35 K (Fig. 3h), reveals that substantial changes occur across the Néel temperature. In particular, the shape of the $\bar{\Gamma}$ -point EDC is modified: a pronounced intensity dip owing to the Dirac point gap at around 0.28 eV seen at 10.5 K (Fig. 3g) is absent at 35 K (Fig. 3h). Nevertheless, the Dirac point gap is present in Fig. 3h, as the EDC fit was only possible using two peaks, corresponding to the Dirac point gap edges. Thus, although the paramagnet–antiferromagnet transition does not lead to closing of the Dirac point gap, the results in Fig. 3e,g,h reveal clearly the strong sensitivity of the TSS to the AFM ordering.

The presence of a Dirac point gap in the paramagnetic phase has been previously reported for the surface states of magnetically doped topological insulators^{13–15}. In the case of the $\text{Bi}_{2-x}\text{Mn}_x\text{Se}_3$ (0001) surface, resonant scattering processes owing to impurity states in the fundamental bandgap were suggested as a possible reason¹⁵. To check whether similar effects take place in our MnBi_2Te_4 samples, we performed resonant photoemission measurements at the Mn 3p–3d edge. The results, shown in Fig. 3i, reveal no resonant features and hence no Mn 3d states around the Dirac point gap, whereupon we discard such a mechanism for the Dirac point gap opening in MnBi_2Te_4 .

To gain a deeper insight into the electronic properties of MnBi_2Te_4 both below and above T_N , we performed electron spin resonance (ESR) measurements (Fig. 4a; D samples). In MnBi_2Te_4 , the ESR linewidth $\Delta H(T)$ shows surprising anisotropic behaviour above approximately 50 K. Namely, the so-called Korringa slope $b = d[\Delta H(T)]/dT$ of the linear $\Delta H(T)$ dependence for $H \perp (0001)$ is three times larger than that for $H \parallel (0001)$ (Fig. 4a), suggesting spatially anisotropic fluctuations of the local magnetic moments owing to their coupling to conduction electrons^{16,17}. In MnBi_2Te_4 , $b_{\perp} \gg b_{\parallel}$, which indicates a relatively short transversal relaxation time τ_2 of the local magnetic moments for the out-of-plane static field geometry and, thus, fast fluctuations of the Mn spins in the (0001) plane. By contrast, in the case of the in-plane field geometry, where the τ_2 relaxation time characterizes Mn spin fluctuations perpendicular to the (0001) plane, the fluctuations appear to be considerably slower, as reflected in the smaller value of b_{\parallel} . Consequently, on a timescale of the photoexcitation process in ARPES (about 10^{-15} s), which is much shorter

than the Mn spin relaxation time (about 10^{-10} s), the probability of finding a Mn spin oriented perpendicularly to (0001) may be larger than for other directions, which could effectively generate an instantaneous out-of-plane field acting on the TSS electrons even in the absence of long-range order. Indeed, our room-temperature (300 K) spin-resolved ARPES measurements reveal the out-of-plane spin components both at the $\bar{\Gamma}$ -point (Fig. 4b) and at finite k_{\parallel} (see Extended Data Fig. 10). Moreover, the calculated paramagnons indicate the presence of ferromagnetic correlations within individual septuple-layer blocks above T_N (Fig. 4a, right inset). Herewith, the ferromagnetically correlated moments tend to point perpendicular to the (0001) plane, which is consistent with the spin-ARPES observations, revealing out-of-plane spin polarization near the Fermi level. We thus conclude that the instantaneous out-of-plane spin polarization observed in our experiments could be responsible for the persistence of the Dirac point gap observed by ARPES in the paramagnetic state of MnBi_2Te_4 .

Our experimental and theoretical results establish MnBi_2Te_4 as an AFM topological insulator. Although in our experiment MnBi_2Te_4 is n doped (which is typical for Bi-based bulk topological insulator crystals), it is this n doping that enables the measurement of both the TSS and the bulk energy gap with ARPES, which probes only the occupied states. A common strategy in synthesizing truly insulating topological insulator crystals is Sb-doping of the Bi sublattice of tetradymite-like compounds^{2,18}, which is expected to work for MnBi_2Te_4 as well. Note that such a tuning of composition is supposed not to affect its interlayer AFM ordering¹⁹. On the other hand, recent progress in the growth of topological insulator films by molecular beam epitaxy²⁰ gives rise to the prospect that nearly-charge-neutral MnBi_2Te_4 can be fabricated.

An AFM topological insulator with the type of antiferromagnetic order established here for MnBi_2Te_4 represents an ideal platform for observing the half-integer quantum Hall effect⁶ ($\sigma_{xy} = e^2/(2h)$; where σ_{xy} and e are the Hall conductivity and electron charge, respectively), which may facilitate experimental confirmation of so-called $\theta = \pi$ quantized magnetoelectric coupling. A material showing this effect is known as an axion insulator, which until now has been sought for in magnetically doped sandwich-like topological insulator heterostructures^{21,22}. Unfortunately, these have been found to show superparamagnetic behaviour^{4,23} that is not seen in MnBi_2Te_4 , which makes it a promising intrinsic axion insulator candidate. Also, the ferromagnetic septuple-layer blocks of MnBi_2Te_4 can be used for the fabrication of topologically nontrivial heterostructures^{24,25}, which are promising for realizing the quantum anomalous Hall effect². Beyond topotronics, another direction of further studies of MnBi_2Te_4 and related materials^{19,26–29} lies within the rapidly growing field of van der Waals magnets^{30,31}. Strongly thickness-dependent properties, expected for van der Waals compounds in the two-dimensional limit, as well as the magnetic degrees of freedom and strong SOC of MnBi_2Te_4 , make it an interesting candidate with which to combine the emerging fields of antiferromagnetic spintronics^{32,33} and layered van der Waals materials^{30,31}.

We have become aware of recent theoretical^{34,35} and experimental^{36–42} studies on MnBi_2Te_4 that confirm our results. In particular, the intrinsic axion insulator state has recently been realized in AFM (that is, below the Néel temperature) MnBi_2Te_4 thin flakes exfoliated from a bulk single crystal⁴¹, providing an additional proof of the AFM topological insulator state in MnBi_2Te_4 below T_N . Additionally, the quantized Hall effect under external magnetic field has been achieved in such flakes^{40–42}. Although the observation of this effect requires the ferromagnetic state of MnBi_2Te_4 , it could not be observed if MnBi_2Te_4 were topologically trivial⁴³.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions

and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1840-9>.

1. Tokura, Y., Yasuda, K. & Tsukazaki, A. Magnetic topological insulators. *Nat. Rev. Phys.* **1**, 126–143 (2019).
2. Chang, C.-Z. et al. Experimental observation of the quantum anomalous Hall effect in a magnetic topological insulator. *Science* **340**, 167–170 (2013).
3. He, Q. L. et al. Chiral Majorana fermion modes in a quantum anomalous Hall insulator–superconductor structure. *Science* **357**, 294–299 (2017).
4. Lachman, E. O. et al. Visualization of superparamagnetic dynamics in magnetic topological insulators. *Sci. Adv.* **1**, e1500740 (2015).
5. Lee, I. et al. Imaging Dirac-mass disorder from magnetic dopant atoms in the ferromagnetic topological insulator $\text{Cr}_x(\text{Bi}_{1-x}\text{Sb}_{0.9})_{2-x}\text{Te}_3$. *Proc. Natl Acad. Sci. USA* **112**, 1316–1321 (2015).
6. Mong, R. S. K., Essin, A. M. & Moore, J. E. Antiferromagnetic topological insulators. *Phys. Rev. B* **81**, 245209 (2010).
7. Qi, X.-L., Hughes, T. L. & Zhang, S.-C. Topological field theory of time-reversal invariant insulators. *Phys. Rev. B* **78**, 195424 (2008).
8. Essin, A. M., Moore, J. E. & Vanderbilt, D. Magnetoelectric polarizability and axion electrodynamics in crystalline insulators. *Phys. Rev. Lett.* **102**, 146805 (2009).
9. Li, R., Wang, J., Qi, X.-L. & Zhang, S.-C. Dynamical axion field in topological magnetic insulators. *Nat. Phys.* **6**, 284–288 (2010).
10. Wang, J., Lian, B. & Zhang, S.-C. Dynamical axion field in a magnetic topological insulator superlattice. *Phys. Rev. B* **93**, 045115 (2016).
11. Lee, D. S. et al. Crystal structure, properties and nanostructuring of a new layered chalcogenide semiconductor, Bi_2MnTe_4 . *CrystEngComm* **15**, 5532–5538 (2013).
12. Hasan, M. Z. & Kane, C. L. Topological insulators. *Rev. Mod. Phys.* **82**, 3045–3067 (2010).
13. Chen, Y. L. et al. Massive Dirac fermion on the surface of a magnetically doped topological insulator. *Science* **329**, 659–662 (2010).
14. Xu, S.-Y. et al. Hedgehog spin texture and Berry's phase tuning in a magnetic topological insulator. *Nat. Phys.* **8**, 616–622 (2012).
15. Sánchez-Barriga, J. et al. Nonmagnetic band gap at the Dirac point of the magnetic topological insulator $(\text{Bi}_{1-x}\text{Mn}_x)_2\text{Se}_3$. *Nat. Commun.* **7**, 10559 (2016).
16. Vaknin, D., Davidov, D., Zevin, V. & Selig, H. Anisotropy and two-dimensional effects in the ESR properties of OsF_6 -graphite intercalation compounds. *Phys. Rev. B* **35**, 6423–6431 (1987).
17. Vithayathil, J. P., MacLaughlin, D. E., Koster, E., Williams, D. L. & Bucher, E. Spin fluctuations and anisotropic nuclear relaxation in single-crystal UPt_3 . *Phys. Rev. B* **44**, 4705–4708 (1991).
18. Zhang, J. et al. Band structure engineering in $(\text{Bi}_{1-x}\text{Sb}_x)_2\text{Te}_3$ ternary topological insulators. *Nat. Commun.* **2**, 574 (2011).
19. Ereemeev, S. V., Otrokov, M. M. & Chulkov, E. V. Competing rhombohedral and monoclinic crystal structures in MnPn_2Ch_4 compounds: an ab-initio study. *J. Alloys Compd.* **709**, 172–178 (2017).
20. Wu, L. et al. Quantized Faraday and Kerr rotation and axion electrodynamics of a 3D topological insulator. *Science* **354**, 1124–1127 (2016).
21. Mogi, M. et al. A magnetic heterostructure of topological insulators as a candidate for an axion insulator. *Nat. Mater.* **16**, 516–521 (2017).
22. Xiao, D. et al. Realization of the axion insulator state in quantum anomalous Hall sandwich heterostructures. *Phys. Rev. Lett.* **120**, 056801 (2018).
23. Krieger, J. A. et al. Spectroscopic perspective on the interplay between electronic and magnetic properties of magnetically doped topological insulators. *Phys. Rev. B* **96**, 184402 (2017).
24. Otrokov, M. M. et al. Magnetic extension as an efficient method for realizing the quantum anomalous Hall state in topological insulators. *JETP Lett.* **105**, 297–302 (2017).
25. Otrokov, M. M. et al. Highly-ordered wide bandgap materials for quantized anomalous Hall and magnetoelectric effects. *2D Mater.* **4**, 025082 (2017).
26. Hirahara, T. et al. Large-gap magnetic topological heterostructure formed by subsurface incorporation of a ferromagnetic layer. *Nano Lett.* **17**, 3493–3500 (2017).
27. Hagmann, J. A. et al. Molecular beam epitaxy growth and structure of self-assembled $\text{Bi}_2\text{Se}_3/\text{Bi}_2\text{MnSe}_4$ multilayer heterostructures. *New J. Phys.* **19**, 085002 (2017).
28. Rienks, E. D. L. et al. Large magnetic gap at the Dirac point in a Mn-induced Bi_2Te_3 heterostructure. Preprint at <https://arxiv.org/abs/1810.06238> (2018).
29. Ereemeev, S. V., Otrokov, M. M. & Chulkov, E. V. New universal type of interface in the magnetic insulator/topological insulator heterostructures. *Nano Lett.* **18**, 6521–6529 (2018).
30. Gong, C. et al. Discovery of intrinsic ferromagnetism in two-dimensional van der Waals crystals. *Nature* **546**, 265–269 (2017).
31. Huang, B. et al. Layer-dependent ferromagnetism in a van der Waals crystal down to the monolayer limit. *Nature* **546**, 270–273 (2017).
32. Baltz, V. et al. Antiferromagnetic spintronics. *Rev. Mod. Phys.* **90**, 015005 (2018).
33. Šmejkal, L., Mokrousov, Y., Yan, B. & MacDonald, A. H. Topological antiferromagnetic spintronics. *Nat. Phys.* **14**, 242–251 (2018).
34. Zhang, D. et al. Topological axion states in magnetic insulator MnBi_2Te_4 with the quantized magnetoelectric effect. *Phys. Rev. Lett.* **122**, 206401 (2019).
35. Li, J. et al. Intrinsic magnetic topological insulators in van der Waals layered MnBi_2Te_4 -family materials. *Sci. Adv.* **5**, eaaw5685 (2019).
36. Gong, Y. et al. Experimental realization of an intrinsic magnetic topological insulator. *Chin. Phys. Lett.* **36**, 076801 (2019).
37. Lee, S. H. et al. Spin scattering and noncollinear spin structure-induced intrinsic anomalous Hall effect in antiferromagnetic topological insulator MnBi_2Te_4 . *Phys. Rev. Res.* **1**, 012011(R) (2019).
38. Yan, J.-Q. et al. Crystal growth and magnetic structure of MnBi_2Te_4 . *Phys. Rev. Mater.* **3**, 064202 (2019).
39. Chen, B. et al. Intrinsic magnetic topological insulator phases in the Sb doped MnBi_2Te_4 bulks and thin flakes. *Nat. Commun.* **10**, 4469 (2019).

40. Deng, Y. et al. Magnetic-field-induced quantized anomalous Hall effect in intrinsic magnetic topological insulator MnBi_2Te_4 . Preprint at <https://arxiv.org/abs/1904.11468> (2019).
41. Liu, C. et al. Quantum phase transition from axion insulator to Chern insulator in MnBi_2Te_4 . Preprint at <https://arxiv.org/abs/1905.00715> (2019).
42. Ge, J. et al. High-Chern-number and high-temperature quantum Hall effect without Landau levels. Preprint at <https://arxiv.org/abs/1907.09947> (2019).
43. Otrokov, M. M. et al. Unique thickness-dependent properties of the van der Waals interlayer antiferromagnet MnBi_2Te_4 films. *Phys. Rev. Lett.* **122**, 107202 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

¹Centro de Física de Materiales (CFM-MPC), Centro Mixto CSIC-UPV/EHU, San Sebastián, Spain. ²IKERBASQUE, Basque Foundation for Science, Bilbao, Spain. ³Donostia International Physics Center (DIPC), San Sebastián, Spain. ⁴Saint Petersburg State University, Saint

Petersburg, Russia. ⁵Experimentelle Physik VII, Universität Würzburg, Würzburg, Germany. ⁶Faculty of Chemistry and Food Chemistry, Technische Universität Dresden, Dresden, Germany. ⁷Institute of Physics, Azerbaijan National Academy of Sciences, Baku, Azerbaijan. ⁸Azerbaijan State Oil and Industry University, Baku, Azerbaijan. ⁹Institute for Solid State Research, Leibniz IFW Dresden, Dresden, Germany. ¹⁰Departamento de Física de Materiales UPV/EHU, San Sebastián, Spain. ¹¹Institut für Theoretische Physik, Johannes Kepler Universität, Linz, Austria. ¹²Tomsk State University, Tomsk, Russia. ¹³Institute of Strength Physics and Materials Science, Russian Academy of Sciences, Tomsk, Russia. ¹⁴Elektronenspeicherring BESSY II, Helmholtz-Zentrum Berlin für Materialien und Energie, Berlin, Germany. ¹⁵Institute of Catalysis and Inorganic Chemistry, Azerbaijan National Academy of Science, Baku, Azerbaijan. ¹⁶Institute of Solid State Physics, Russian Academy of Sciences, Chernogolovka, Russia. ¹⁷Faculty of Physics, Technische Universität Dresden, Dresden, Germany. ¹⁸Hiroshima Synchrotron Radiation Center, Hiroshima University, Higashi-Hiroshima, Japan. ¹⁹Department of Physical Sciences, Graduate School of Science, Hiroshima University, Higashi-Hiroshima, Japan. ²⁰Elettra Sincrotrone Trieste, Trieste, Italy. ²¹Advanced Light Source, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. ²²Max-Planck-Institut für Mikrostrukturphysik, Halle, Germany. *e-mail: mikhail.otrokov@gmail.com; evguenivladimirovich.tchoukov@ehu.eus

Methods

Electronic structure and total-energy calculations

Electronic structure calculations were carried out using density functional theory using the projector augmented-wave method⁴⁴, implemented in the Vienna Ab initio Simulation Package (VASP)^{45,46}. The exchange–correlation energy was treated using the generalized gradient approximation⁴⁷. The Hamiltonian contained scalar relativistic corrections and the SOC was taken into account by the second variation method⁴⁸. To describe the van der Waals interactions we made use of the DFT-D2⁴⁹ and the DFT-D3^{50,51} approaches, which gave similar results. The energy cutoff for the plane-wave expansion was set to 270 eV. All structural optimizations were performed using a conjugate-gradient algorithm and a force-tolerance criterion for convergence of 0.01 eV Å⁻¹. SOC was always included when performing relaxations.

The Mn 3*d* states were treated employing the GGA+*U* approach⁵² within the Dudarev scheme⁵³. The $U_{\text{eff}} = U - J$ value (where U and J are the effective on-site Coulomb and exchange interaction parameters, respectively) for the Mn 3*d* states was chosen to be equal to 5.34 eV, as in previous work^{24–26}. Using this U_{eff} we found a good agreement with the HSE06 functional^{54–56} in the fundamental bandgap and binding energy of the Mn 3*d* states of bulk MnBi₂Te₄. Also, we find a very good agreement of the calculated bandgap for the single MnBi₂Te₄ septuple-layer block (0.32 eV)⁴³ with the measured one (0.35 eV)³⁶. Moreover, the indirect character of the gap is correctly reproduced as well. Note that the GGA itself unsatisfactorily describes the bandgap of a single septuple-layer MnBi₂Te₄ film both in terms of the character (it yields direct gap) and size (43 meV); using GGA+*U* improves the description of the *p*–*d* hybridization in the system. Further testing was performed to check the stability of the results against U_{eff} variation. Namely, the bulk crystal structure was fully optimized for different U_{eff} values (3 eV, 4 eV and 5.34 eV) and then the magnetic ordering was studied. It was found that the AFM ground state does not change upon such variations of U_{eff} and the crystal structure. The MnBi₂Te₄ magnetic anisotropy was found to be stable against these variations as well.

To model the interlayer AFM structure in MnBi₂Te₄, we used a rhombohedral cell with 14 atoms. These calculations were performed with a three-dimensional Brillouin zone sampled by a 9 × 9 × 9 *k*-point grid.

The magnetic anisotropy energy, $E_a = E_{\text{diff}} + E_d$, was calculated taking into account the total energy differences of various magnetization directions, $E_{\text{diff}} = E_{\text{in-plane}} - E_{\text{out-of-plane}}$, and the energy of the classical dipole–dipole interaction, E_d . To calculate E_{diff} , the energies for three inequivalent magnetization directions (Cartesian *x*, *y* (in-plane) and *z* (out-of-plane)) were calculated and E_{diff} was determined to be the difference $E_{\text{in-plane}} - E_z$, where $E_{\text{in-plane}}$ is the energy of the most energetically favourable in-plane direction of the magnetization. A *k*-mesh of 2,197 points was chosen, and the total energies were calculated self-consistently for all considered directions. The calculations of E_{diff} were done for the interlayer AFM state in the cell containing 14 atoms (Fig. 1d). The energy convergence criterion was set to 10⁻⁷ eV, giving a well converged E_{diff} (up to a few tenths of a millielectronvolt) and excluding ‘accidental’ convergence. A cutoff radius of at least 20 μm was used to calculate E_d .

The \mathbb{Z}_2 invariant for the three-dimensional AFM topological insulator was calculated according to ref. ⁵⁷. When spatial inversion symmetry is present in the system, as it is in the MnBi₂Te₄ case, the following \mathbb{Z}_2 -invariant ζ_0 can be defined:

$$(-1)^{\zeta_0} = \prod_{\mathbf{k}_{\text{inv}} \in \text{B-TRIM}, n \in \text{occ}/2} \zeta_n(\mathbf{k}_{\text{inv}})$$

where ζ_n is the parity of the *n*th occupied (occ) band and B-TRIM are special momenta \mathbf{k}_{inv} in which each level is doubly degenerate, see ref. ⁵⁷ for further details.

The MnBi₂Te₄ semi-infinite surface was simulated within a model of repeating films separated by a vacuum gap of a minimum of 10 Å. A 56-atomic-layers-thick slab was used, which corresponds to eight septuple layers. The interlayer distances were optimized for the topmost septuple-layer block of each surface. Both the structural optimizations and static electronic structure calculations were performed using a *k*-point grid of 11 × 11 × 1 in the two-dimensional Brillouin zone.

Exchange-coupling constants calculations

For the equilibrium structures obtained with VASP, we calculated the Heisenberg exchange-coupling constants J_{ij} also from first principles, this time using the full-potential linearized augmented plane waves (FLAPW) formalism⁵⁸ as implemented in FLEUR⁵⁹. We took the GGA+*U* approach^{60,61} under the fully localized limit⁶². For the self-consistent FLAPW basis set in the MnBi₂Te₄ compound we chose a dense 22 × 22 × 22 Monkhorst–Pack *k*-point sampling of the first Brillouin zone and a cutoff of 3.4 hartree (1 hartree = 27.21 eV; R_{∞} is the Rydberg constant). The density and potential expansions were cut at 10.4 hartree. Locally, muffin-tin-sphere radii values of 2.74 atomic units for Mn and 2.81 atomic units for Bi and Te atoms were used, and the partial wavefunctions were expanded up to cutoffs of the orbital angular momentum, *l* = 8. Mn, Bi, and Te contribute 4*s*3*d*, 5*s*5*p* and 6*s*6*p* valence electrons, respectively. We verified that these settings accurately reproduce the bandstructures obtained with VASP, both with and without SOC effects. The ferromagnetic versus AFM ordering energy differences are also in agreement.

The J_{ij} constants were extracted by Fourier inversion of the magnon energy dispersion for the MnBi₂Te₄ primitive cell^{59,63,64}, neglecting SOC. These dispersion energies, calculated in the force theorem approach, correspond to a constrained set of non-collinear spin configurations characterized by the magnon *q* vectors of a 13 × 13 × 13 grid^{59,65}. The reference self-consistent electron wavefunctions were obtained with a 12 × 12 × 12 *k*-grid. These grids ensured that magnon energies converged below 0.1 meV and enabled us to add up to 150 neighbouring atoms to the Fourier analysis, ensuring accurate J_{ij} values.

Paramagnons calculations

Paramagnetic fluctuations were calculated within a first-principles approach based on the coherent potential approximation^{66–69}. The disordered local moment method was used to take ensemble averages over the orientational configurations of the local moments⁷⁰. Paramagnons are calculated considering the response of the disordered local moment paramagnetic state to the application of an external, site-dependent, magnetic field⁷¹.

Ab-initio-based tight-binding calculations

Ab-initio-based tight-binding calculations were performed using the VASP package with the Wannier90 interface^{72,73}. The Wannier basis chosen consisted of six spinor *p*-type orbitals $|p_x^{\uparrow}\rangle, |p_y^{\uparrow}\rangle, |p_z^{\uparrow}\rangle, |p_x^{\downarrow}\rangle, |p_y^{\downarrow}\rangle, |p_z^{\downarrow}\rangle$ of Bi and Te. The surface electronic band structure was calculated within the semi-infinite medium Green’s function approach^{74,75}.

Monte Carlo simulations

The Monte Carlo simulations were based on a classical Heisenberg Hamiltonian that includes the magnetic anisotropy energy E_a

$$\mathcal{H} = -\frac{1}{2} \sum_{ij} J_{ij} \mathbf{e}_i \cdot \mathbf{e}_j + \sum_i E_a (\mathbf{e}_i^z)^2$$

where the magnetic moments at site *i* and *j* are described by unit vectors \mathbf{e}_i and \mathbf{e}_j , respectively, and the magnetic coupling constants J_{ij} are determined by ab initio calculations as described above. The term of the magnetic anisotropy takes into account only the *z* component of the magnetic moment. For the simulation of the bulk system we created a cluster with periodic boundary conditions. Therein, the cluster size was varied with the primitive unit cell repeated *N* times in all directions with

$N=12, 14, \dots, 20$. Varying the system size in this way enabled us to avoid finite size effects. We tentatively started the calculation at a temperature of 60 K and reduced T stepwise until 0.001 K. At each temperature step, the thermal equilibrium was reached after 40,000 Monte Carlo steps (chosen after comparing results for 20,000 to 100,000 steps for all systems). The same number of steps was then used to derive the observables after reaching thermal equilibrium. One Monte Carlo step represents the creation of a new random spin direction, which is either accepted or not depending on the energy difference and the current temperature. The critical temperature was finally obtained from the peak in the magnetic susceptibility, which represented the clearest identification criterion.

Crystal growth

Dresden samples (D samples). High-quality bulk single crystals of MnBi_2Te_4 were grown from the melt by slow cooling of a 1:1 mixture of the binary compounds Bi_2Te_3 and $\alpha\text{-MnTe}$. The binaries were synthesized by mechanical pre-activation and annealing of stoichiometric mixtures of the elements. Crystal size and quality were controlled via different cooling rates within a narrow temperature interval at around 600 °C and varying annealing times. Further details of crystal-growth optimization are reported elsewhere⁷⁶. Single-crystal X-ray diffraction was measured on a four-circle CCD diffractometer (Kappa APEX II, Bruker) with a graphite(002)-monochromator and a CCD-detector at $T=296(2)$ K. Mo- $\text{K}\alpha$ radiation ($\lambda=71.073$ pm) was used. A numerical absorption correction based on an optimized crystal description⁷⁷ was applied, and the initial structure solution was performed in the JANA2006⁷⁸ software. The structure was refined in the SHELXL program against the experimentally observed squared structure factors F_o^2 (ref. ^{79,80}). The crystal structure has been deposited in the joint Cambridge Crystallographic Data Centre/FIZ Karlsruhe under depository number CSD-1867581. The structure refinement yields some degree of statistical cation disorder in the Mn and Bi positions in contrast to an earlier reported ordered model¹¹. However, Mn/Bi antisite defects in two fully occupied cation positions do not lead to a superstructure ordering or change of translational symmetry. Energy dispersive X-ray spectra (EDS) were collected using a Silicon Drift X-MaxN (Oxford Instruments) detector at an acceleration voltage of 20 kV and an accumulation time of 100 s. The analysis was performed using the $P/B\text{-ZAF}$ standardless method (where Z = atomic number correction factor, A = absorption correction factor, F = fluorescence factor and P/B = peak-to-background model). Energy dispersive X-ray spectra reproducibly yielded a stoichiometric composition, ruling out the possibility of large compositional variations in our samples.

Baku samples (B samples). The bulk ingot of the Baku sample was grown from the melt with a non-stoichiometric composition using the vertical Bridgman method. The pre-synthesized polycrystalline sample was evacuated in a conical-bottom quartz ampoule sealed under vacuum better than 10^{-4} Pa. In order to avoid any reaction during the melting process between the Mn content of the sample and the silica container, the inside wall of the ampoule was coated with graphite by thermal decomposition of acetone in an oxygen-poor environment. The ampoule was held in the ‘hot’ zone (about 680 °C) of a two-zone tube furnace of the MnBi_2Te_4 Bridgman crystal-growth system for 8 h to achieve a complete homogenization of the melts. Then, it moved from the upper (hot) zone to the bottom (cold) zone with a required rate of 0.7 mm h^{-1} . Consequently, we obtained a bulk ingot with average dimensions of 3 cm in length and 0.8 cm in diameter. Further details are reported elsewhere⁸¹. The as-grown ingot was checked by X-ray diffraction measurements and was found to consist of several single crystalline blocks. With the aid of X-ray diffraction data, high-quality single crystalline pieces were isolated from different parts of the as-grown ingot for further measurements.

Magnetic measurements

The magnetic measurements as a function of temperature and magnetic field were performed on a stack of single crystals of MnBi_2Te_4 (D samples) using a SQUID (superconducting quantum interference device) VSM (vibrating-sample magnetometer) (Quantum Design). The temperature-dependent magnetization measurements were acquired in external magnetic fields of 0.02 T and 1 T for both zero-field-cooled and field-cooled-warming conditions. A thorough background subtraction was performed for all curves.

Part of the magnetic measurements were carried out at the Center for Diagnostics of Materials for Medicine, Pharmacology and Nanoelectronics of the SPbU Science Park using a SQUID magnetometer with a helium cryostat (Quantum Design). The measurements were carried out in a pull mode in terms of temperature and magnetic field. The applied magnetic field was perpendicular to the (0001) sample surface.

Resistivity measurements

Resistivity measurements were done with a standard four-probe ac technique using a low-frequency ($f \approx 20$ Hz) lock-in amplifier. Contacts were attached with conducting graphite paste. The measurements were carried out in a temperature-variable cryostat at different values of magnetic field up to 8 T, generated by a superconducting solenoid and directed along the normal to the (0001) sample surface.

Temperature- and field-dependent resistivity measurements were performed on B samples (Extended Data Fig. 3a). The metallic-like behaviour characteristic of the presence of free carriers is observed at $H=0$ as the resistivity ρ increases with rising temperature. This is consistent with the results of the Hall-effect measurements yielding the n-type conductivity of these samples (Extended Data Fig. 3b). A well defined kink at 25.4 K indicates a magnetic transition in agreement with the magnetization studies and Monte Carlo simulations. In a series of measurements under an external field $H \perp (0001)$, the kink shifts to lower temperatures as the field increases from 1 T to 3 T. Above the critical field (around 3 T to 4 T), the $\rho(T)$ slope is much steeper below T_N , which could be related to the observed spin-flop in the $M(H)$ curve (Fig. 2e).

ESR measurements

ESR experiments were performed with a commercial X-band ESR-spectrometer (EMX, Bruker) operating at a microwave frequency of 9.6 GHz and providing magnetic fields up to 0.9 T. It is equipped with a He gas flow cryostat (Oxford Instruments) and a goniometer allowing temperature- and angular-dependent measurements between 4 K and 300 K.

The temperature dependence of the resonance field H_{res} of the ESR signal of Mn^{2+} ions in MnBi_2Te_4 is measured for the out-of-plane $H \perp (0001)$ and in-plane $H \parallel (0001)$ orientations of the static magnetic field H . At high temperatures H_{res} is almost isotropic within the error bars of ± 80 Oe, and the corresponding spectroscopic g -factor is very close to the spin-only value $g_s = 2$, as expected for the Mn^{2+} ($3d^5$; spin angular momentum $S = 5/2$; orbital angular momentum $L = 0$) ion⁸². However, below $T \approx 50$ K, H_{res} becomes anisotropic and, in particular, the resonance line rapidly shifts to smaller fields for the out-of-plane geometry—indicating the onset of quasi-static short-range magnetic correlations in MnBi_2Te_4 well above the AFM phase transition at T_N . By further lowering the temperature, the linewidth experiences critical broadening due to the slowing down of the spin fluctuations in the vicinity⁸³ of T_N (Fig. 4a). Finally, upon entering the AFM-ordered state the intensity of the ESR signal rapidly decreases, owing to the shifting of the spectral weight to the AFM collective resonance modes, which typically occur at much higher frequencies than the paramagnetic resonance⁸⁴, and thus out of the frequency range of our ESR setup. Observation of quasi-static short-range magnetic correlations in the ESR experiment is consistent with the strong spin fluctuation-driven

spin scattering above T_N found in a previous magneto-transport study³⁷ of MnBi_2Te_4 .

ARPES measurements

The ARPES experiments were carried out at the BaDElPh beamline⁸⁵ of the Elettra synchrotron in Trieste (Italy) and BL-1 of the Hiroshima synchrotron radiation centre (Japan) using p polarization of the synchrotron radiation and laser^{86,87}. The photoemission spectra were collected on freshly cleaved surfaces. The base pressure during the experiments was better than 1×10^{-10} mbar. Some of the ARPES experiments were also carried out at the resource centre Physical Methods of Surface Investigation (PMSI) at the research park of Saint Petersburg State University.

Additional $h\nu$ -dependent experiments on D samples (data shown in Extended Data Fig. 7) were performed at the MAESTRO endstation of the Advanced Light Source facility.

Dichroic ARPES measurements

The linear dichroism ARPES measurements on D samples (Extended Data Fig. 8) were performed at the MAESTRO endstation of the Advanced Light Source facility. This measurement enabled us to explore the influence of the AFM state on the wavefunction of the TSS. With p-polarized light incident in the x - z plane the linear dichroism (LD) in the photoelectron intensity (I) can be defined as^{88–91}:

$$\text{LD}(k_x, k_y, E) = I(k_x, k_y, E) - I(-k_x, k_y, E)$$

The linear dichroism is the intensity asymmetry relative to the y - z plane. In the present experiment the y - z plane is a crystalline mirror plane. Therefore, in the absence of a magnetization, the linear dichroism is induced by the mirror-symmetry breaking of the light electric field vector $\mathbf{E} = (\varepsilon_x, 0, \varepsilon_z)$. The intensities along $\pm k_x$ can be written as $I(\pm k_x) = |T_z \pm T_x|^2$, where T_z and T_x are the photoemission matrix elements of the electric field components ε_z and ε_x between the photoelectron final state Φ_f and the initial state ψ_i at a given k_x (ref. ⁸⁸). We find $\text{LD} = 4\Re(T_z^* T_x)$, where \Re denotes the real part and the asterisk indicates the complex conjugate, implying that a change of either matrix element will also change the linear dichroism. In particular, a change of the TSS wavefunction across T_N will manifest in the linear dichroism because the TSS enters the matrix elements as the initial-state wavefunction ψ_i .

This is precisely what we observe in Extended Data Fig. 8a–c: at temperatures above T_N a linear dichroism is already present, but it considerably increases below T_N . This effect is understood as the effect of the AFM order on ψ_i . In the present case, as described in the text, T_N marks the transition between the AFM state and a paramagnetic state with anisotropic fluctuations. Our linear-dichroism measurements provide strong evidence for an effect of the AFM order on the TSS wavefunction.

Our theoretical calculations provide additional insight into the origin of the linear dichroism. The crystal structure of the (0001) surface of MnBi_2Te_4 has three mirror planes along the Γ - \bar{M} directions of the two-dimensional Brillouin zone. In the nonmagnetic case, the presence of these mirror planes dictates that the out-of-plane spin components are zero for the Γ - \bar{M} directions, along which the spins are locked exclusively within the surface plane. However, for the Γ - \bar{K} directions the s_z components are allowed by symmetry and can therefore coexist with the in-plane spins. As can be seen in a previous work²⁴, in the case of an isostructural compound GeBi_2Te_4 , ab initio calculations reveal the presence (absence) of the out-of-plane spins for the Γ - \bar{K} (Γ - \bar{M}) directions. At GeBi_2Te_4 (0001), the out-of-plane spin components are especially pronounced away from the fundamental bandgap energy region, where the TSS coincides with the bulk states in energy (but not in k_{\parallel}). By analogy, the appearance of the out-of-plane spin components along Γ - \bar{K} can be expected for the TSS of MnBi_2Te_4 (0001). Such a spin texture is required to respect the time-reversal symmetry if the AFM exchange field inherent to the material is artificially set to zero. In this case, the

sign of s_z will change upon changing $+k_{\parallel}$ (right branch) to $-k_{\parallel}$ (left branch). When the time-reversal symmetry is broken and each MnBi_2Te_4 septuple layer is ferromagnetically ordered, the right and left branches of the TSS interact differently with the Zeeman field provided by the Mn layer of the topmost septuple-layer block (Extended Data Fig. 8d). This is similar to the so-called exchange+Rashba effect^{92,93}. In such cases, a dispersion asymmetry is created—that is, $E(+k_{\parallel}) \neq E(-k_{\parallel})$ —which is indeed seen for the MnBi_2Te_4 TSS (Extended Data Fig. 8d, e), resulting in a sizeable asymmetry in the orbital composition between its right and left branches (Extended Data Fig. 8e, f). Note that such asymmetries are not found in our calculations for the Γ - \bar{M} directions because of the symmetry constraints on the presence of s_z components in the case of time-reversal-symmetry preservation. In the time-reversal symmetry breaking case, the s_z components appear along the Γ - \bar{M} directions, which also breaks the mirror symmetry at the MnBi_2Te_4 (0001) surface.

Temperature-dependent laser ARPES measurements

To study the Dirac cone state response to the onset of the AFM order, we carried out the temperature-dependent high-resolution laser ARPES measurements for the MnBi_2Te_4 (0001) surface. In Extended Data Fig. 9a, the raw EDC profiles at the $\bar{\Gamma}$ -point at 10.5 K and 35 K are presented (as in Fig. 3g, h), clearly showing the increase of the intensities of both the upper and lower parts of the Dirac cone state below the Néel temperature. Before studying the Dirac cone intensity variation systematically, we analysed the intensity of the first two bulk conduction-band states (those shown in the Extended Data Fig. 5c). No temperature dependence of the integral intensity of these bulk states was observed (Extended Data Fig. 9b) and therefore the EDC spectra require no specific temperature-dependent normalization. Thus, the intensity temperature dependence was obtained using the raw ARPES data. The results are shown in Extended Data Fig. 9c, where the $\bar{\Gamma}$ -point EDC profiles, acquired with a temperature step of around 0.9 K, are plotted versus temperature. Below about 24–25 K, the increase in intensity in the $\Delta E = [0.2 \text{ eV}, 0.3 \text{ eV}]$ energy window, containing the Dirac cone state at the $\bar{\Gamma}$ -point, is clearly seen. Note that these temperature-dependent ARPES spectra were measured with the equal acquisition time and the constant laser photon flux for each temperature point in Extended Data Fig. 9c. In order to obtain a deeper insight into this intensity behaviour, we integrate it within the energy window ΔE and plot as a function of temperature in Extended Data Fig. 9d (the same data are presented in Fig. 3e of the manuscript). We see that although the Dirac cone intensity is roughly constant above $T_N \approx 24$ K, below the Néel temperature it starts to grow, and at 10 K it increases by a factor of 1.4–1.5 compared to its value at 35 K.

We note that during the ARPES measurements the spectra acquisition time is limited, which is due to a finite ‘lifetime’ of the clean surface of the sample even under the ultrahigh-vacuum conditions. This lifetime is comparable to the time needed to acquire several high-quality ARPES maps (such as shown in Fig. 3g, h), that are suitable for a reliable comparative EDC spectral analysis (fitting). However, within the lifetime of the clean surface of the sample, such a long acquisition time per temperature point is unreachable during the systematic temperature-dependent measurements that in our case are made with more than 50 temperature points, that is, $\Delta T \approx 0.9$ K and two sweep directions (10 K \rightarrow 35 K \rightarrow 10 K). Although the measured EDCs shown in Extended Data Fig. 9c are clearly meaningful, their spectral decomposition (fitting), similar to that shown in Extended Data Fig. 9a, is not unambiguous. Treated without any normalization, these raw data clearly show a strong change of intensity across the Néel temperature.

As can be seen from the EDC analysis (Extended Data Fig. 9a), the intensity increase observed is largely a result of the increase in intensity of the gapped Dirac cone peaks; the signal between those peaks (that is, within the Dirac point gap) is practically the same for 10.5 K and 35 K. Therefore, at 35 K the intensity inside the Dirac point gap increases

relative to those of the Dirac cone peaks, indicative of an increase in the spectral weight inside the gap. The temperature-dependent change of the line shape around the Dirac point gap can be nicely demonstrated by the behaviour of the second derivative ($d^2N(E)/dE^2$), which is highly sensitive to the curvature of the spectrum $N(E)$ (see Extended Data Fig. 9e). Indeed, at low temperatures we see a pronounced feature of the second derivative around the binding energy of 0.25 eV, signalling the presence of the Dirac point gap, although above the Néel temperature this feature is substantially weaker. These changes demonstrate the gapped state mitigation across the antiferromagnet–paramagnet transition.

Spin-resolved ARPES measurements

Spin-resolved ARPES measurements were performed at the RBL-2 end-station of BESSY II in Berlin (Germany) using a hemispherical analyser (Scienta R4000) and a photon energy of 6 eV. The photon beam was generated using the fourth harmonic of a custom-made femtosecond-laser system coupled to an ultrafast amplifier operating at a repetition rate of 100 kHz. The spin-resolved spectra were acquired with a three-dimensional Mott-type spin detector operated at 26 kV. We used the vertical linearly polarized light incident on the sample under an angle of 45° with respect to the surface normal. The experimental geometry is given in ref.⁹⁴. The energy and angular resolutions of the spin-resolved measurements were 30 meV and 1.5° (corresponding to 0.02 Å⁻¹), respectively.

Resonant photoemission measurements

Resonant photoemission data were acquired at the HR-ARPES branch of the I05 beamline at the Diamond Light Source. The measurements were conducted at a base temperature of $T = 10$ K with beam spot size and energy resolution of $A_{\text{spot}} \approx 50 \times 50 \mu\text{m}^2$ and $E \approx 20$ meV, respectively. The difference between on- and off-resonance spectra for the Mn 3p–3d transition corresponds directly to the density of the Mn 3d states. A photon energy series was conducted to determine suitable transition energies. The corresponding angle integrated spectra of on-resonance ($h\nu = 51$ eV) and off-resonance ($h\nu = 47$ eV) conditions can be seen in Fig. 3i.

XMCD measurements

Surface-sensitive XMCD measurements (total electron yield mode)⁹⁵ on D samples were performed at the HECTOR end-station of the BOREAS beamline at the ALBA synchrotron radiation facility⁹⁶. The spot size and the resolving power of the supplied photon beam were $A_{\text{spot}} < 200 \times 200 \mu\text{m}^2$ and $E/\Delta E > 9,000$, respectively. Measurements were performed at the Mn L_{2,3} edges at a temperature of 2 K, that is, well below $T_N \approx 24$ K.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request. The crystal structure is available in the joint Cambridge Crystallographic Data Centre/FIZ Karlsruhe (<https://www.ccdc.cam.ac.uk/structures/>) under the deposit number CSD-1867581.

44. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).

45. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).

46. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).

47. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).

48. Koelling, D. D. & Harmon, B. N. A technique for relativistic spin-polarised calculations. *J. Phys. C* **10**, 3107 (1977).

49. Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **27**, 1787–1799 (2006).

50. Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H–Pu. *J. Chem. Phys.* **132**, 154104 (2010).

51. Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **32**, 1456–1465 (2011).

52. Anisimov, V. I., Zaanen, J. & Andersen, O. K. Band theory and Mott insulators: Hubbard U instead of Stoner I. *Phys. Rev. B* **44**, 943–954 (1991).

53. Dudarev, S. L., Botton, G. A., Savrasov, S. Y., Humphreys, C. J. & Sutton, A. P. Electron-energy-loss spectra and the structural stability of nickel oxide: An LSDA+U study. *Phys. Rev. B* **57**, 1505–1509 (1998).

54. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **38**, 3098 (1988).

55. Perdew, J. P., Ernzerhof, M. & Burke, K. Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys.* **105**, 9982–9985 (1996).

56. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **118**, 8207–8215 (2003).

57. Fang, C., Gilbert, M. J. & Bernevig, B. A. Topological insulators with commensurate antiferromagnetism. *Phys. Rev. B* **88**, 085406 (2013).

58. Wimmer, E., Krakauer, H., Weinert, M. & Freeman, A. J. Full-potential self-consistent linearized-augmented-plane-wave-method for calculating the electronic structure of molecules and surfaces: O₂ molecule. *Phys. Rev. B* **24**, 864–875 (1981).

59. FLEUR <http://www.flapw.de>, version fleur.26e (2017).

60. Anisimov, V. I., Aryasetiawan, F. & Lichtenstein, A. I. First-principles calculations of the electronic structure and spectra of strongly correlated systems: the LDA+U method. *J. Phys. Condens. Matter* **9**, 767 (1997).

61. Shick, A. B., Liechtenstein, A. I. & Pickett, W. E. Implementation of the LDA+U method using the full-potential linearized augmented plane-wave basis. *Phys. Rev. B* **60**, 10763–10769 (1999).

62. Anisimov, V. I., Solov'yev, I. V., Korotin, M. A., Czyżyk, M. T. & Sawatzky, G. A. Density-functional theory and NiO photoemission spectra. *Phys. Rev. B* **48**, 16929–16934 (1993).

63. Sandratskii, L. M. & Bruno, P. Exchange interactions and Curie temperature in (Ga,Mn)As. *Phys. Rev. B* **66**, 134435 (2002).

64. Ležaić, M., Mavropoulos, P., Enkovaara, J., Bihlmayer, G. & Blügel, S. Thermal collapse of spin polarization in half-metallic ferromagnets. *Phys. Rev. Lett.* **97**, 026404 (2006).

65. Kurz, P., Förster, F., Nordström, L., Bihlmayer, G. & Blügel, S. Ab initio treatment of noncollinear magnets with the full-potential linearized augmented plane wave method. *Phys. Rev. B* **69**, 024415 (2004).

66. Soven, P. Coherent-potential model of substitutional disordered alloys. *Phys. Rev.* **156**, 809 (1967).

67. Gyorffy, B. L. Coherent-potential approximation for a nonoverlapping-muffin-tin-potential model of random substitutional alloys. *Phys. Rev. B* **5**, 2382 (1972).

68. Lüders, M., Ernst, A., Temmerman, W. M., Szotek, Z. & Durham, P. J. Ab initio angle-resolved photoemission in multiple-scattering formulation. *J. Phys. Condens. Matter* **13**, 8587 (2001).

69. Geilhufe, M. et al. Numerical solution of the relativistic single-site scattering problem for the Coulomb and the Mathieu potential. *J. Phys. Condens. Matter* **27**, 435202 (2015).

70. Gyorffy, B. L., Pindor, A. J., Staunton, J., Stocks, G. M. & Winter, H. A first-principles theory of ferromagnetic phase transitions in metals. *J. Phys. F* **15**, 1337 (1985).

71. Staunton, J., Gyorffy, B. L., Pindor, A. J., Stocks, G. M. & Winter, H. Electronic structure of metallic ferromagnets above the Curie temperature. *J. Phys. F* **15**, 1387 (1985).

72. Marzari, N. & Vanderbilt, D. Maximally localized generalized Wannier functions for composite energy bands. *Phys. Rev. B* **56**, 12847 (1997).

73. Mostofi, A. A. et al. wannier90: A tool for obtaining maximally-localized Wannier functions. *Comput. Phys. Commun.* **178**, 685–699 (2008).

74. Lopez Sancho, M. P., Lopez Sancho, J. M., Sancho, J. M. L. & Rubio, J. Highly convergent schemes for the calculation of bulk and surface Green functions. *J. Phys. F* **15**, 851–858 (1985).

75. Henk, J. & Schattke, W. A subroutine package for computing Green's functions of relaxed surfaces by the renormalization method. *Comput. Phys. Commun.* **77**, 69–83 (1993).

76. Zeugner, A. et al. Chemical aspects of the candidate antiferromagnetic topological insulator MnBi₂Te₄. *Chem. Mater.* **31**, 2795–2806 (2019).

77. X-Shape, Crystal Optimization for Numerical Absorption Correction Program Version 2.12.2, <https://www.stoe.com/product/software-x-area/> (Stoe & Cie, 2009).

78. Petricek, V., Dusek, M. & Palatinus, L. Jana2006 <http://jana.fzu.cz> (Institute of Physics, 2011).

79. Sheldrick, G. M. SHELXL Version 2014/7 <https://shelx.uni-goettingen.de> (Georg-August-Universität Göttingen, 2014).

80. Sheldrick, G. M. A short history of SHELX. *Acta Crystallogr. A* **64**, 112–122 (2008).

81. Aliev, Z. S. et al. Novel ternary layered manganese bismuth tellurides of the MnTe–Bi₂Te₃ system: synthesis and crystal structure. *J. Alloys Compd.* **789**, 443–450 (2019).

82. Abragam, A. & Bleaney, B. *Electron Paramagnetic Resonance of Transition Ions* (Oxford University Press, 2012).

83. Benner, H. & Boucher, J. In *Magnetic Properties of Layered Transition Metal Compounds* 323–378 (Kluwer, 1990).

84. Turov, E. A. *Physical Properties of Magnetically Ordered Crystals* (Academic Press, 1965).

85. Petaccia, L. et al. BaD EIPh: a 4-m normal-incidence monochromator beamline at Elettra. *Nucl. Instrum. Methods Phys. Res. A* **606**, 780–784 (2009).

86. Iwasawa, H. et al. Rotatable high-resolution ARPES system for tunable linear-polarization geometry. *J. Synchrotron Radiat.* **24**, 836–841 (2017).

87. Iwasawa, H. et al. Development of laser-based scanning μ-ARPES system with ultimate energy and momentum resolutions. *Ultramicroscopy* **182**, 85–91 (2017).

88. Bentmann, H. et al. Strong linear dichroism in spin-polarized photoemission from spin-orbit-coupled surface states. *Phys. Rev. Lett.* **119**, 106401 (2017).

89. Chernov, S. V. et al. Anomalous d-like surface resonances on Mo(110) analyzed by time-of-flight momentum microscopy. *Ultramicroscopy* **159**, 453–463 (2015).

90. Tusche, C. et al. Multi-MHz time-of-flight electronic bandstructure imaging of graphene on Ir(111). *Appl. Phys. Lett.* **108**, 261602 (2016).

91. Schönhenne, G. et al. Spin-filtered time-of-flight *k*-space microscopy of Ir towards the complete photoemission experiment. *Ultramicroscopy* **183**, 19–29 (2017).
92. Krupin, O. et al. Rashba effect at magnetic metal surfaces. *Phys. Rev. B* **71**, 201403 (2005).
93. Rybkin, A. G. et al. Magneto-spin-orbit graphene: interplay between exchange and spin-orbit couplings. *Nano Lett.* **18**, 1564–1574 (2018).
94. Sánchez-Barriga, J. et al. Subpicosecond spin dynamics of excited states in the topological insulator Bi₂Te₃. *Phys. Rev. B* **95**, 125405 (2017).
95. Abbate, M. et al. Probing depth of soft X-ray absorption spectroscopy measured in total-electron-yield mode. *Surf. Interface Anal.* **18**, 65–69 (1992).
96. Barla, A. et al. Design and performance of BOREAS, the beamline for resonant X-ray absorption and scattering experiments at the ALBA synchrotron light source. *J. Synchrotron Radiat.* **23**, 1507–1517 (2016).

Acknowledgements M.M.O. and E.V.C. thank A. Arnau and J. I. Cerdá for discussions. We acknowledge support by the Basque Departamento de Educacion, UPV/EHU (grant number IT-756-13), the Spanish Ministerio de Economia y Competitividad (MINECO grant number FIS2016-75862-P), and the Academic D.I. Mendeleev Fund Program of Tomsk State University (project number 8.1.01.2018). Support from the Saint Petersburg State University grant for scientific investigations (grant ID 40990069), the Russian Science Foundation (grants number 18-12-00062 for part of the photoemission measurements and 18-12-00169 for part of the calculations of topological invariants and tight-binding bandstructure calculations), the Russian Foundation for Basic Research (grant number 18-52-06009), and the Science Development Foundation under the President of the Republic of Azerbaijan (grant number EIF-BGM-4-RFTF-1/2017-21/04/1-M-02) is also acknowledged. M.M.O. acknowledges support by the Diputación Foral de Gipuzkoa (project number 2018-CIEN-000025-01). I.I.K. and A.M.S. acknowledge partial support from the CERIC-ERIC consortium for the stay at the Elettra synchrotron. The ARPES measurements at HISOR were performed with the approval of the Proposal Assessing Committee (proposal numbers 18AG020, 18BU005). The support of the German Research Foundation (DFG) is acknowledged by A.U.B.W., A.I. and B.B. within Collaborative Research Center 1143 (SFB 1143, project ID 247310070); by A.Z., A.E. and A.I. within Special Priority Program 1666 Topological Insulators; by H.B. and F.R. within Collaborative Research Center 1170; and by A.Z. and A.I. within the ERANET-Chemistry Program (RU 776/15-1). H.B., A.U.B.W., A.A., V.K., B.B., F.R. and A.I. acknowledge financial support from the DFG through the Würzburg-Dresden Cluster of Excellence on Complexity and Topology in Quantum Matter – ct.qmat (EXC 2147, project ID 39085490). A.E. acknowledges support by the OeAD grant numbers HR 07/2018 and PL 03/2018. This work was

also supported by the Fundamental Research Program of the State Academies of Sciences, line of research III.23. A.K. was financially supported by KAKENHI number 17H06138 and 18H03683. I.P.R. acknowledges support by the Ministry of Education and Science of the Russian Federation within the framework of the governmental program Megagrants (state task number 3.8895.2017/P220). E.V.C. acknowledges financial support by the Gobierno Vasco-UPV/EHU project (IT1246-19). S.K. acknowledges financial support from an Overseas Postdoctoral Fellowship, SERB-India (OPDF award number SB/OS/PDF-060/2015-16). J.S.-B. acknowledges financial support from the Impuls-und Vernetzungsfonds der Helmholtz-Gemeinschaft under grant number HRSF-0067 (Helmholtz-Russia Joint Research Group). The calculations were performed in Donostia International Physics Center, in the research park of Saint Petersburg State University Computing Center (<http://cc.spbu.ru>), and at Tomsk State University.

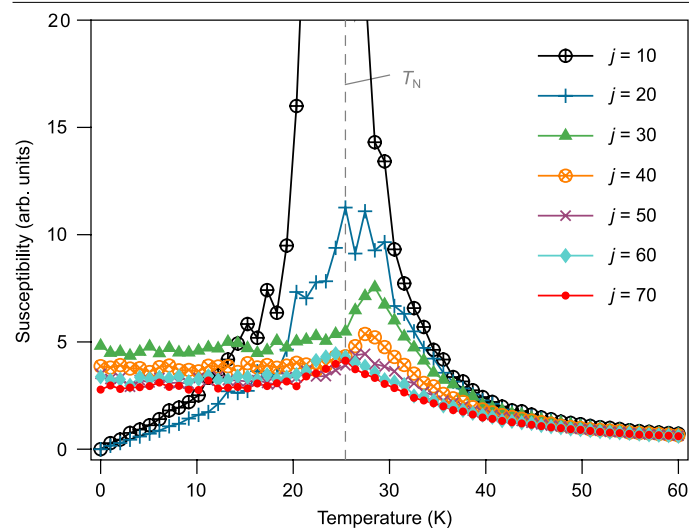
Author contributions The bandstructure calculations were performed by M.M.O., M.B.-R., S.V.E. and A.Yu.V. The exchange-coupling constants calculations were performed by M.B.-R., Yu.M.K., M.M.O. and A.E. The paramagnons calculations were performed by A.E. The magnetic anisotropy studies were performed by M.M.O. The Monte Carlo simulations were performed by M.H. The topological invariant calculations were done by S.V.E. Tight-binding calculations were performed by I.P.R. and V.M.K. Crystals were grown by A.Z., A.I., Z.S.A. and M.B.B. X-ray diffraction measurements and structure determination were performed by A.Z. and I.R.A. The resistivity and Hall measurements, as well as contact preparation, were done by N.T.M., N.A.A. and V.N.Z. XMCD and resonant photoemission experiments were performed by R.C.V., T.R.F.P., C.H.M., K.K., S.S. and H.B. Magnetization experiments and their analysis were mainly performed by S.G., B.B. and A.U.B.W. with contributions by A.V.K. ARPES measurements were done by I.I.K., D.E., A.M.S., E.F.S., S.K., A.K., L.P., G.D.S., R.C.V., K.K., M.Ü., S.M. and H.B. The analysis of the ARPES data was done by I.I.K., D.E., R.C.V. and H.B. Spin-ARPES measurements were performed by I.I.K., D.E., A.M.S., F.F. and J.S.-B. ESR measurements were done by A.A. and V.K. The project was planned by M.M.O., A.I., H.B., A.M.S., N.T.M., F.R., P.M.E. and E.V.C. The supervision of the project was executed by E.V.C. All authors contributed to the discussion and manuscript editing. The paper was written by M.M.O. with contributions from A.I., H.B., V.K. and A.U.B.W.

Competing interests The authors declare no competing interests.

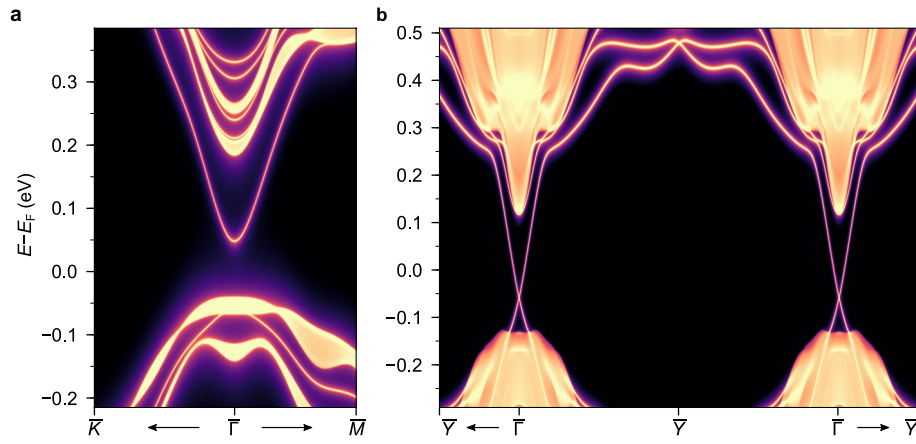
Additional information

Correspondence and requests for materials should be addressed to M.M.O. or E.V.C.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

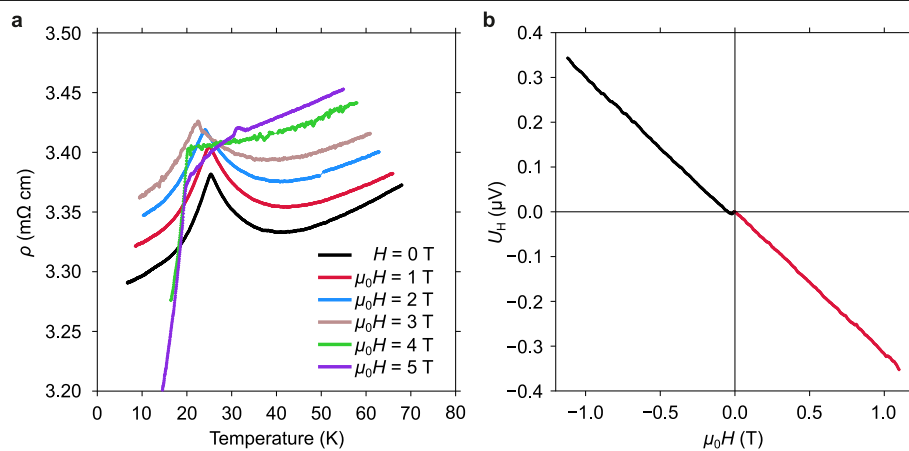


Extended Data Fig. 1 | Monte Carlo simulations for bulk MnBi_2Te_4 . The temperature-dependent magnetic susceptibility of bulk MnBi_2Te_4 calculated for various numbers of magnetic shells j up to which the exchange-coupling constants J_{ij} were considered in the classical Heisenberg Hamiltonian. In increments of 10, results for 10–70 shells were calculated. The vertical dashed line shows the final Néel temperature of 25.4 K, estimated from the calculation for 70 shells. Note that the simulations revealed the onset of the AFM ground state only above 20 shells.



Extended Data Fig. 2 | Electronic structure of the S-breaking and S-preserving surfaces of MnBi_2Te_4 . **a, b,** Surface electronic bandstructure of MnBi_2Te_4 calculated for the (0001) (S-breaking; **a**), and (10 $\bar{1}$ 1) (S-preserving;

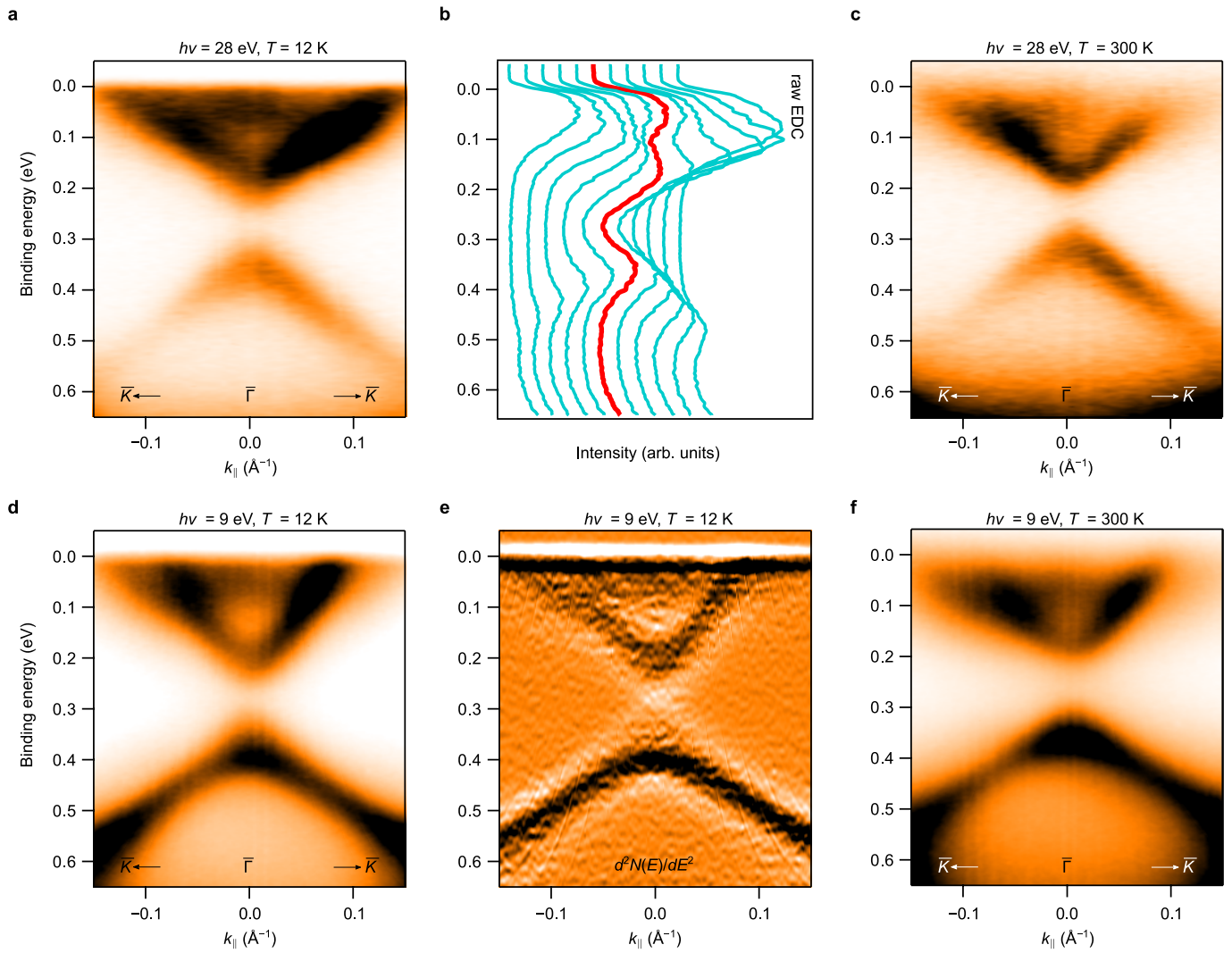
b) terminations using the ab-initio-based tight-binding approach. The regions with a continuous spectrum correspond to the three-dimensional bulk states projected onto a two-dimensional Brillouin zone.



Extended Data Fig. 3 | Resistivity and Hall measurements of bulk MnBi_2Te_4 .

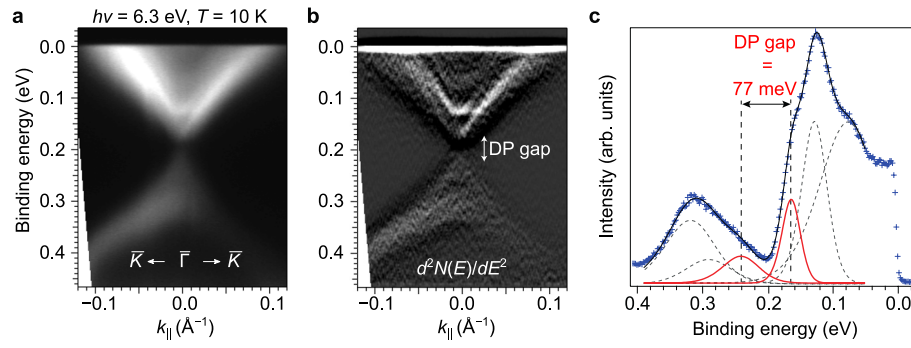
a, Temperature- and field-dependent resistivity data. **b**, Hall voltage as a function of the applied magnetic field for MnBi_2Te_4 at 5 K. The Hall-effect measurements unambiguously indicate n-type conductivity for our MnBi_2Te_4

samples that show a negative Hall voltage for positive values of the applied magnetic field. The estimated electron concentration and Hall mobility are around $2 \times 10^{19} \text{ cm}^{-3}$ and $100 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$, respectively. The measurements were performed on B samples.



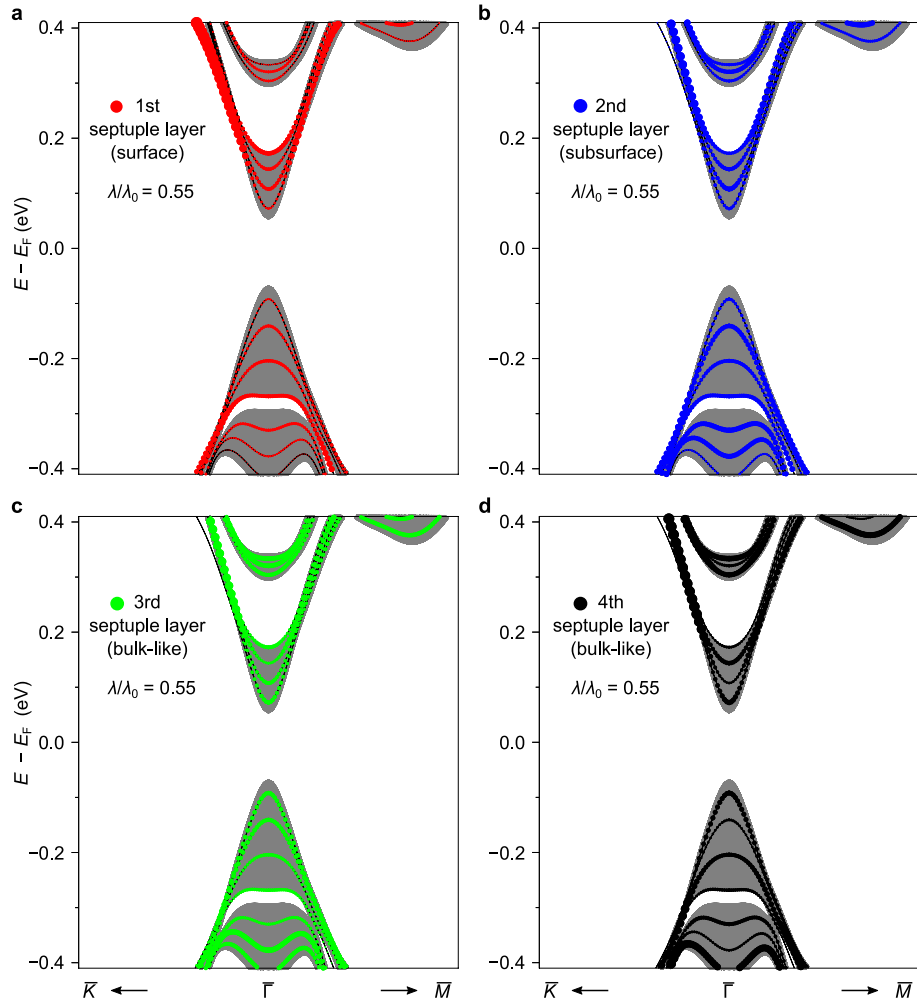
Extended Data Fig. 4 | D-sample ARPES. **a**, Dispersion of MnBi₂Te₄(0001) measured at 12 K with a photon energy of 28 eV. **b**, EDC representation of the data in **a**. The red curve marks the EDC at the $\bar{\Gamma}$ -point. **c**, ARPES image acquired at 300 K ($h\nu = 28$ eV). **d**, Dispersion of MnBi₂Te₄(0001) measured at a

temperature of 12 K with a photon energy of 9 eV (which is more bulk sensitive). **e**, The second derivative $d^2N(E)/dE^2$ of the data in **d**. **f**, ARPES image acquired at 300 K ($h\nu = 9$ eV). The measurements were performed on D samples.



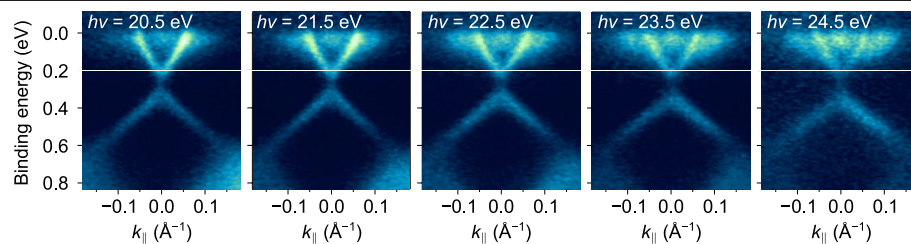
Extended Data Fig. 5 | Laser-based ARPES. **a**, ARPES map of MnBi_2Te_4 taken at 10 K with a photon energy of 6.3 eV. In this case, the photoelectrons have a kinetic energy of about 1.5 eV and, subsequently, a large mean free path in the sample, corresponding to a high bulk sensitivity of this experiment. **b**, The second derivative $d^2N(E)/dE^2$ of the data in **a**. **c**, Fitting results for the EDC

spectrum at the $\bar{\Gamma}$ -point. The raw data, the resulting fitting curve and its decomposition with Voigt peaks are shown by blue symbols, a black solid line and the grey dashed and red solid lines, respectively. Red (grey) lines indicate the peaks attributed to the gapped Dirac cone state (bulk bands). The measurements were performed on B samples.



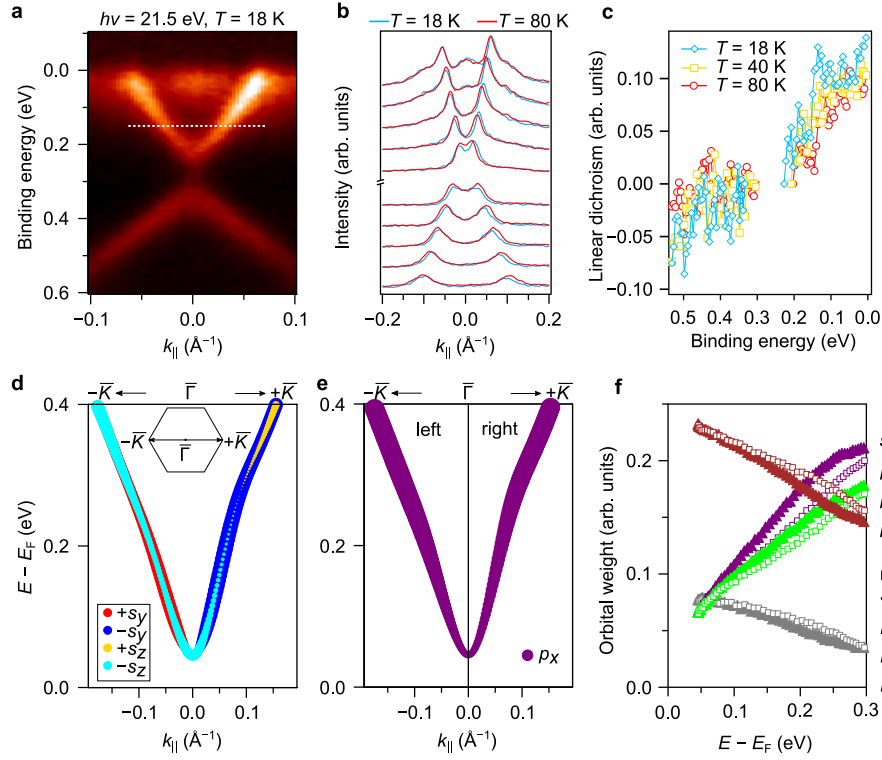
Extended Data Fig. 6 | Surface electronic structure of MnBi_2Te_4 in the artificial topologically trivial phase. Septuple-layer resolved (0001) surface electronic structure of MnBi_2Te_4 calculated for the SOC constant value $\lambda = 0.55\lambda_0$. The size of the colour circles that comprise the data reflects the state localization in a particular septuple-layer block of the eight-septuple-layer-thick slab. **a**, First septuple layer (that is, the surface layer; red). **b**, Second septuple layer (subsurface; blue). **c**, Third septuple layer (bulk-like; green).

d, Fourth septuple layer (bulk-like; black). The grey areas correspond to the bulk bandstructure projected onto the surface Brillouin zone. We see that near the Γ -point there are (1) no surface states in the bulk bandgap and (2) no resonance states near the bandgap edges. The first quantum-well states of both the valence and conduction bands are strongly localized in the inner parts of the slab.



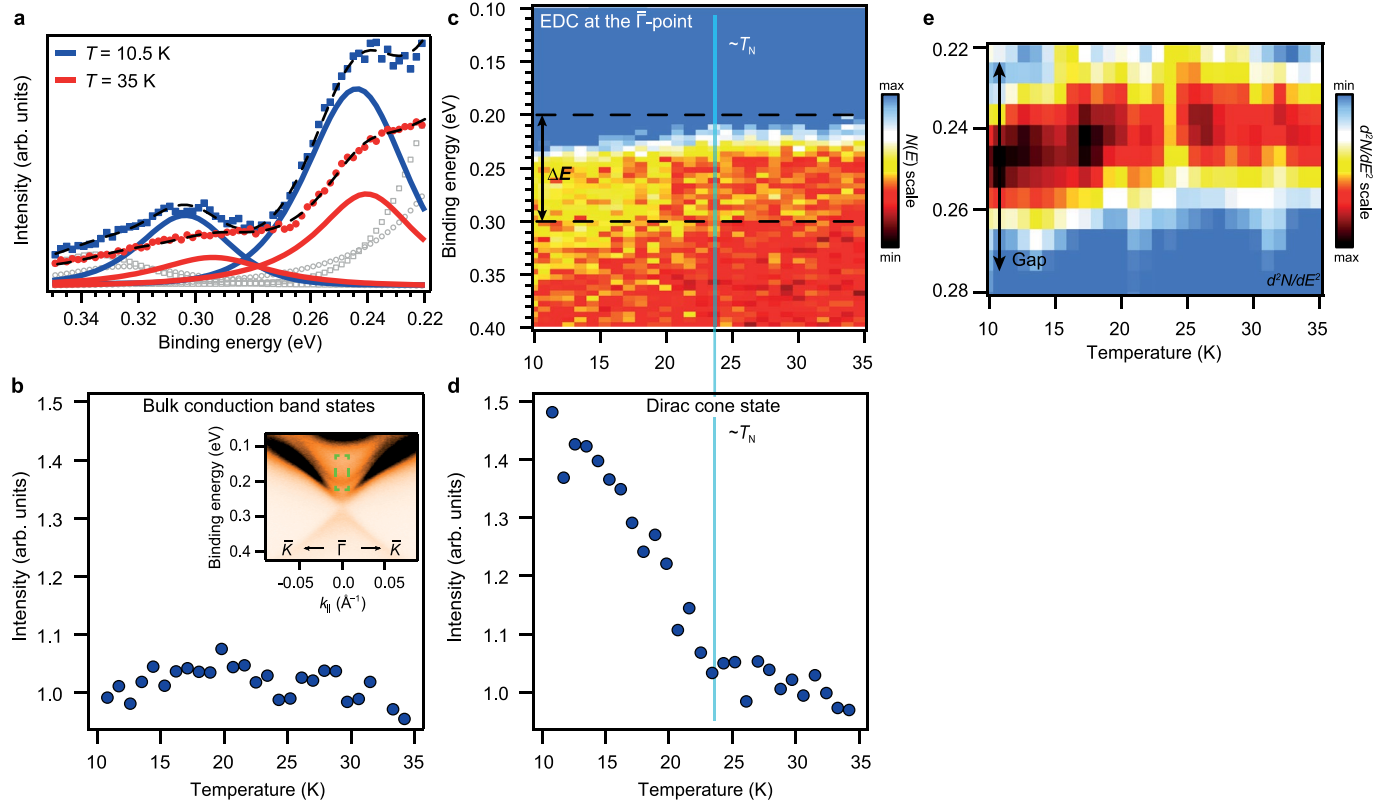
Extended Data Fig. 7 | Photon-energy-dependent ARPES data. Photon-energy-dependent ARPES data measured near the Brillouin zone centre along the \bar{K} – $\bar{\Gamma}$ – \bar{K} direction at a temperature of 18 K. Absence of any $h\nu$ dependence

confirms the surface-state character of the upper cone. The measurements were performed on D samples.



Extended Data Fig. 8 | Temperature-dependent linear dichroism in the Dirac cone photoemission intensity. **a**, Dispersion of MnBi_2Te_4 (0001) measured at 18 K with a photon energy of 21.5 eV and p-polarized light along the $\bar{K}-\Gamma-\bar{K}$ direction. **b**, Momentum distribution curves representation of the data acquired at 18 K (blue) and 80 K (red). **c**, Linear dichroism ($I_{\text{right}} - I_{\text{left}}$), where I_{right} and I_{left} are the intensities of the right and left branches of the upper and lower cone corresponding to positive and negative k_{\parallel} , respectively. The measurements were performed on D samples. **d**, Upper part of the MnBi_2Te_4 (0001) gapped Dirac cone as calculated ab initio. The size of the coloured circles reflects the value and sign of the spin vector Cartesian projections, with red (blue) corresponding to the positive (negative) s_y

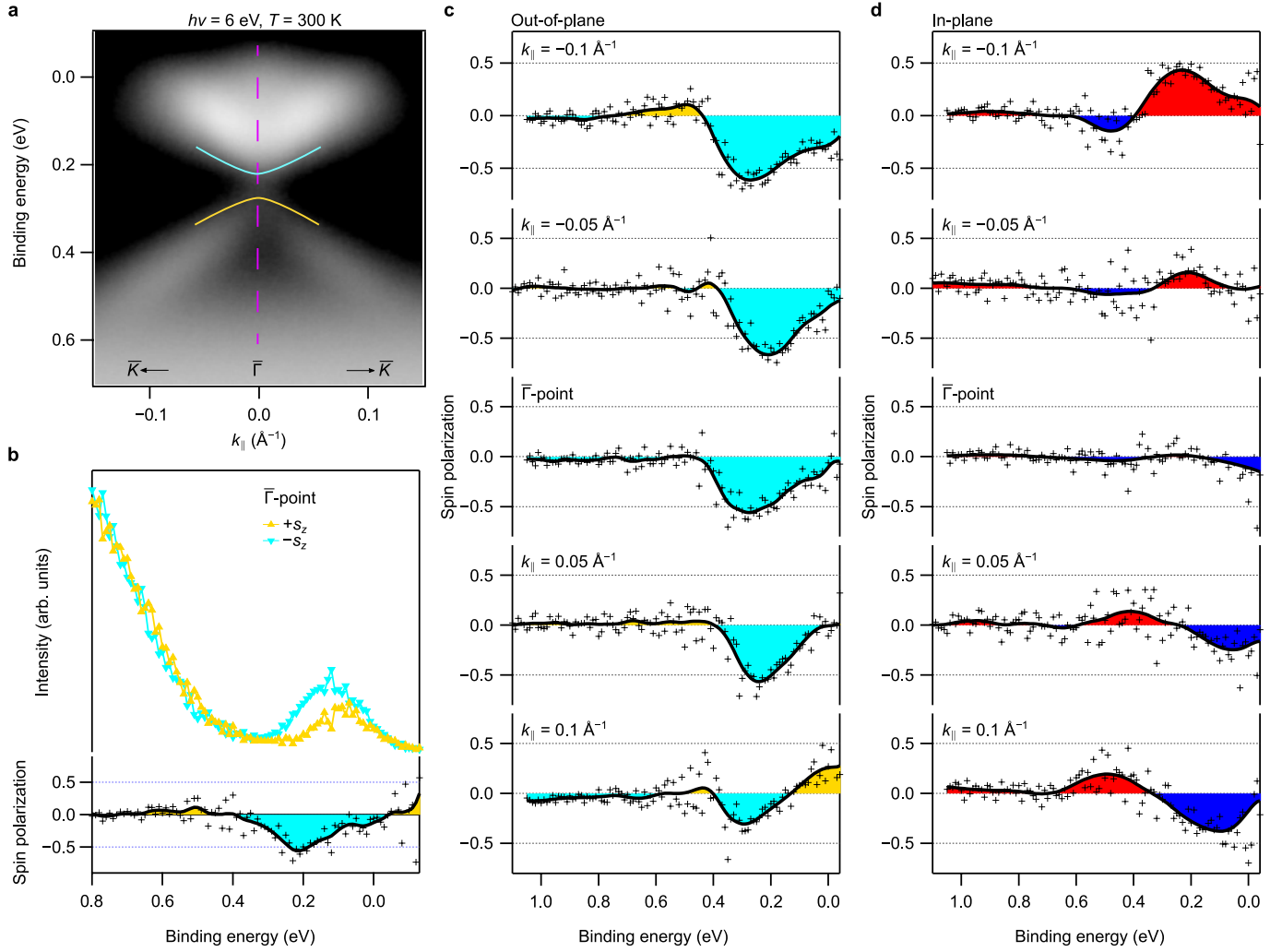
components (perpendicular to k_z), and yellow (cyan) to the out-of-plane components $+s_z$ ($-s_z$). **e**, As in **d**, but with the size of the purple circles reflecting the weight of the p_x orbitals of all Bi and Te atoms of the topmost septuple-layer block at each k_{\parallel} . Note that in **d**, **e** the bulk-like bands of the slab are omitted. The magnetic moment of the topmost Mn layer points towards vacuum, but in Fig. 1e and Extended Data Fig. 6 it points in the opposite direction. **f**, The weight of the s , p_x , p_y and p_z orbitals of all Bi and Te atoms of the topmost septuple-layer block for the left (triangles) and right (squares) branches as a function of energy. See Methods for more information on the dichroic ARPES measurements.



Extended Data Fig. 9 | Temperature-dependent laser ARPES measurements.

a, ARPES EDC profiles taken at the Γ -point of MnBi_2Te_4 (0001) at 10.5 K and 35 K. The raw data, resulting fitted curves, and their decompositions with Voigt peaks are shown by the coloured symbols, the black dashed lines, and the coloured lines and grey symbols, respectively. Red and blue lines (red circles and blue squares) indicate the peaks (EDCs) of the Dirac cone state at 35 K and 10.5 K, respectively. The peaks of the bulk bands at 35 K (10.5 K) are shown by grey circles (squares). **b**, Integrated intensity of the first two bulk conduction-band states (those analysed in detail in Extended Data Fig. 5c) as a function of temperature. Inset, The ARPES MnBi_2Te_4 (0001) map measured with a laser photon energy of 6.4 eV and $T = 10.5$ K (as in Fig. 3d). The green rectangle marks the region of the map where the first two bulk conduction-band states are located. The average intensity in the shown temperature interval was set to 1. **c**, EDC profiles, $N(E)$, taken at the Γ -point between 10 K and 35 K with a temperature step $\Delta T \approx 0.9$ K and two sweep directions (10 K \rightarrow 35 K \rightarrow 10 K).

Because the measurements upon heating and cooling reveal essentially the same behaviour, in **c** we show the data averaged over these two sets of the EDC profiles at each temperature point. Note that the data in **a** and the intensity dependencies on temperature in **b–d** were acquired from two different B samples, showing slightly different binding energy of the Dirac point gap centres (0.28 eV and 0.25 eV, respectively). **d**, Intensity integrated within the energy window ΔE marked by the dashed black lines in **c**. The average intensity in the plateau-like region above approximately 24 K was set to 1. ΔE contains both the lower and upper parts of the Dirac cone at the Γ -point and corresponds to the energy interval in which the contribution of the cone is dominant and that of the bulk states is almost negligible. The vertical cyan line approximately shows the start of the intensity increase, which roughly corresponds to $T_N \approx 24$ K for MnBi_2Te_4 . **e**, The second derivative, $d^2N(E)/dE^2$, of the EDC profiles in **c**, shown for a clearer visualization of the Dirac point gap behaviour.



Extended Data Fig. 10 | Spin-resolved ARPES data. **a**, Spin-integrated ARPES spectrum taken at 6 eV photon energy along the \bar{K} - Γ - \bar{K} direction. Yellow and cyan curves show the location of the gapped TSS. **b**, Spin-resolved ARPES spectra taken at the Γ -point with respect to the out-of-plane spin quantization axis. The out-of-plane spin polarization is shown below the corresponding spin-up and spin-down spectra. **c, d**, Measured out-of-plane (**c**) and in-plane (**d**) spin

polarization at different momentum values. The in-plane spin polarization changes its sign with k_{\parallel} , as expected for the TSS. The change of the out-of-plane spin polarization sign at $k_{\parallel} = +0.1 \text{ \AA}^{-1}$ near the Fermi level in **c** (bottom) is discussed in the Methods section 'Dichroic ARPES measurements'. The data in **a, b** and **c, d** were acquired on B and D samples, respectively. The measurements were performed at $T = 300 \text{ K}$.

Large magnetic gap at the Dirac point in $\text{Bi}_2\text{Te}_3/\text{MnBi}_2\text{Te}_4$ heterostructures

<https://doi.org/10.1038/s41586-019-1826-7>

Received: 10 May 2018

Accepted: 18 October 2019

Published online: 18 December 2019

E. D. L. Rienks^{1,2,3,14}, S. Wimmer^{4,14}, J. Sánchez-Barriga^{1,14}, O. Caha^{5,14}, P. S. Mandal^{1,7}, J. Růžička⁵, A. Ney⁴, H. Steiner⁴, V. V. Volobuev^{4,8,13}, H. Groiss⁹, M. Albu¹⁰, G. Kothleitner¹⁰, J. Michalička⁶, S. A. Khan¹¹, J. Minár¹¹, H. Ebert¹², G. Bauer⁴, F. Freyse^{1,7}, A. Varykhalov¹, O. Rader^{1*} & G. Springholz^{4*}

Magnetically doped topological insulators enable the quantum anomalous Hall effect (QAHE), which provides quantized edge states for lossless charge-transport applications^{1–8}. The edge states are hosted by a magnetic energy gap at the Dirac point², but hitherto all attempts to observe this gap directly have been unsuccessful. Observing the gap is considered to be essential to overcoming the limitations of the QAHE, which so far occurs only at temperatures that are one to two orders of magnitude below the ferromagnetic Curie temperature, T_C (ref. ⁸). Here we use low-temperature photoelectron spectroscopy to unambiguously reveal the magnetic gap of Mn-doped Bi_2Te_3 , which displays ferromagnetic out-of-plane spin texture and opens up only below T_C . Surprisingly, our analysis reveals large gap sizes at 1 kelvin of up to 90 millielectronvolts, which is five times larger than theoretically predicted⁹. Using multiscale analysis we show that this enhancement is due to a remarkable structure modification induced by Mn doping: instead of a disordered impurity system, a self-organized alternating sequence of MnBi_2Te_4 septuple and Bi_2Te_3 quintuple layers is formed. This enhances the wavefunction overlap and size of the magnetic gap¹⁰. Mn-doped Bi_2Se_3 (ref. ¹¹) and Mn-doped Sb_2Te_3 form similar heterostructures, but for Bi_2Se_3 only a nonmagnetic gap is formed and the magnetization is in the surface plane. This is explained by the smaller spin–orbit interaction by comparison with Mn-doped Bi_2Te_3 . Our findings provide insights that will be crucial in pushing lossless transport in topological insulators towards room-temperature applications.

The QAHE was first demonstrated in chromium-doped tetradymite topological insulators^{2–5}. Subsequently, replacing chromium with vanadium was a successful strategy for achieving precise quantization with vanishing longitudinal resistance^{6,7}. The effect occurs because of a modification of the band inversion in the ferromagnetic state. Exchange splitting and spin–orbit coupling lead to a release of the inversion of one of the spin sub-bands². This should manifest itself as a magnetic gap that opens up at the Dirac point when the system is cooled below the Curie temperature. So far, however, direct observation of this gap has remained elusive and no clear correlation with ferromagnetism has been established.

Angle-resolved photoemission spectroscopy (ARPES) is the method of choice for the direct observation of the magnetic gap. Nevertheless, the situation has been confusing: large gaps of 0.05–0.2 eV were

first reported for Mn-doped Bi_2Se_3 (refs. ^{12,13}), but were later shown not to be of magnetic origin¹⁴. Such gaps, however, did not appear when magnetic impurities were deposited on the surface of Bi_2Se_3 (refs. ^{14–16}). At low temperatures, a mobility gap of 32 meV was inferred from scanning tunnelling Landau level spectroscopy of V-doped Sb_2Te_3 (ref. ¹⁷), but scanning tunnelling spectroscopy (STS) did not show a gap in this system¹⁷. STS did reveal gaps of 20–100 meV in Cr-doped $(\text{Bi}, \text{Sb})_2\text{Te}_3$ (ref. ¹⁸), but the temperature dependence was not investigated. In fact, a similar gap of around 75 meV was found for Cr-doped Bi_2Se_3 even at room temperature¹⁹. This suggests a nonmagnetic origin for these effects, because the ferromagnetic T_C is well below 50 K in all of these systems.

Interestingly, the configuration of the magnetic dopants is also contradictory. For isovalent magnetic doping, it has been predicted that

¹Helmholtz-Zentrum Berlin für Materialien und Energie, Elektronenspeicherring BESSY II, Berlin, Germany. ²Institut für Festkörperphysik, Technische Universität Dresden, Dresden, Germany.

³Leibniz-Institut für Festkörper- und Werkstoffforschung Dresden, Dresden, Germany. ⁴Institut für Halbleiter- und Festkörperphysik, Johannes Kepler Universität, Linz, Austria. ⁵Department of Condensed Matter Physics, Masaryk University, Brno, Czech Republic. ⁶Central European Institute of Technology, Brno University of Technology, Brno, Czech Republic. ⁷Institut für Physik und Astronomie, Universität Potsdam, Potsdam, Germany. ⁸National Technical University ‘Kharkiv Polytechnic Institute’, Kharkiv, Ukraine. ⁹Christian Doppler Laboratory for Nanoscale Phase Transformations, Zentrum für Oberflächen- und Nanoanalytik, Johannes Kepler Universität, Linz, Austria. ¹⁰Graz Center for Electron Microscopy, Institute of Electron Microscopy and Nanoanalysis, Graz University of Technology, Graz, Austria. ¹¹New Technologies Research Centre, University of West Bohemia, Pilsen, Czech Republic. ¹²Department Chemie, Ludwig-Maximilians-Universität, München, Germany. ¹³Present address: International Research Centre MagTop and Institute of Physics, Polish Academy of Sciences, Warsaw, Poland. ¹⁴These authors contributed equally: E. D. L. Rienks, S. Wimmer, J. Sánchez-Barriga, O. Caha. *e-mail: rader@helmholtz-berlin.de; gunther.springholz@jku.at

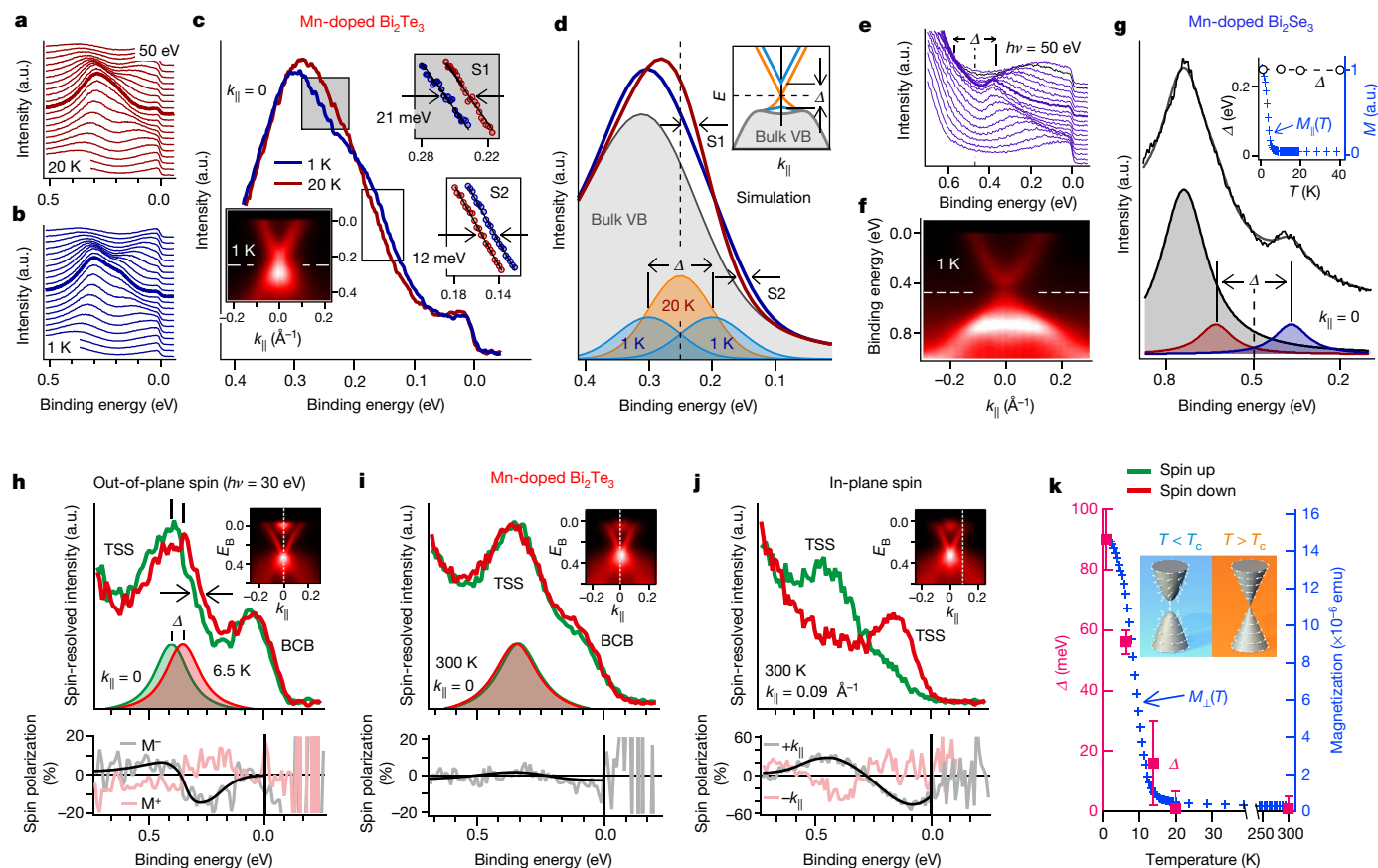


Fig. 1 | Magnetic gap of Mn-doped Bi₂Te₃. **a–d**, ARPES for Bi₂Te₃ with 6% Mn, above and below the Curie temperature (T_C) of around 10 K. The spectra in **c**, **d** and those marked by thick lines in **a**, **b** correspond to the centre of the surface Brillouin zone at $k_{\parallel} = 0$. Line fits in the regions S1 and S2 in **c** yield a splitting of more than 33 meV between 20 K and 1 K; according to the simulations shown in **d**, this splitting corresponds to a magnetic gap, Δ , of 90 ± 10 meV. The inset in **c** shows the energy, E , versus k_{\parallel} momentum map at 1 K. a.u., arbitrary units. VB, valence band. **e–g**, Same analysis for Bi₂Se₃ with 6% Mn and a T_C of 6 K, revealing only a temperature-independent nonmagnetic gap, Δ , that does not correlate with magnetization (see inset). **h, i**, Spin-resolved

ARPES of Bi₂Te₃ with 6% Mn at 6.5 K (**h**) and 300 K (**i**), showing that below T_C the gap at the Dirac point is ferromagnetically spin-split with out-of-plane spin orientation. At 6.5 K, the magnetic gap is $\Delta = 56 \pm 4$ meV. **j**, Away from the Dirac point, the conventional helical in-plane spin texture is measured. The out-of-plane component of the spin polarization reverses with the reversal in magnetization (M) (as shown in the lower part of **h**), and the in-plane spin component reverses with the in-plane wavevector k_{\parallel} (lower part of **j**). **k**, The same temperature dependence is measured for $\Delta(T)$ and the magnetization $M(T)$. The inset shows a sketch of the measured spin textures below and above T_C .

Bi₂Se₃, Bi₂Te₃ and Sb₂Te₃ will form a QAHE state, which should thus occur when Bi or Sb are substituted by Cr or Fe, but not when the substituents are Ti or V, owing to their metallicity². Moreover, nonisovalent magnetic dopants turn out to have surprisingly little effect on carrier concentration: that is, Mn-doped Bi₂Se₃ and Bi₂Te₃ always remain n-type^{14,20}, even though divalent Mn replacing trivalent Bi should act as a strong acceptor.

To resolve these issues, we present a comprehensive study of Mn-doped Bi₂Te₃ and Bi₂Se₃ that unequivocally reveals a large magnetic exchange splitting at the Dirac point of Bi₂Te₃. This splitting vanishes above the Curie temperature, which is clear-cut evidence for its magnetic origin. No increase in the gap size is observed for Mn-doped Bi₂Se₃ at temperatures down to 1 K. Through a multiscale structural analysis, we reveal that the actual lattice structure is very different to the anticipated random impurity system, as Mn doping induces the formation of self-organized heterostructures. This turns out to be crucial for obtaining large magnetic gaps¹⁰.

Bandgap, spin texture and magnetism

Figure 1a–g shows the ARPES dispersions of Mn-doped Bi₂Te₃ and Bi₂Se₃, measured above and below the ferromagnetic phase transition ($T_C = 10$ K and 6 K, respectively). For Mn-doped Bi₂Te₃, the photoemission

spectrum recorded at $h\nu = 50$ eV at the centre of the surface Brillouin zone shows an intensity maximum at a binding energy of 0.3 eV from the bulk valence band, while the Dirac point of the topological surface state (TSS) contributes a smaller peak at around 0.2 eV. On cooling from 20 K through T_C down to 1 K, the low energy flank of the peak develops a pronounced shoulder, forming a plateau at around 0.2 eV (Fig. 1a–c). Assuming that the single component for the topological surface state at 20 K becomes split into two equally intense components at 1 K (Fig. 1d), we arrive at a gap, Δ , of 90 ± 10 meV at low temperature (see Methods section ‘ARPES’ and Extended Data Fig. 1). Because T_C is 10 K in this sample, this proves the magnetic origin of this gap. This is the central result of our study.

To confirm the magnetic origin, we carried out spin-resolved ARPES. For the spectra shown in Fig. 1h–j, we chose a photon energy of $h\nu = 30$ eV at which bulk transitions are strong from the bulk conduction band (BCB) but overlap much less with the Dirac point of the TSS. At 6.5 K, the spectrum at the Dirac point—measured in remanence after field cooling (M)—is clearly spin-polarized, with spin orientation perpendicular to the surface and spin split by $\Delta = 56 \pm 4$ meV (Fig. 1h and Extended Data Fig. 2). Because the temperature is closer to T_C , this value is smaller than that derived at 1 K. Subsequent measurement at room temperature (Fig. 1i) shows that the spin polarization has completely disappeared, whereas subsequent cooling in an oppositely

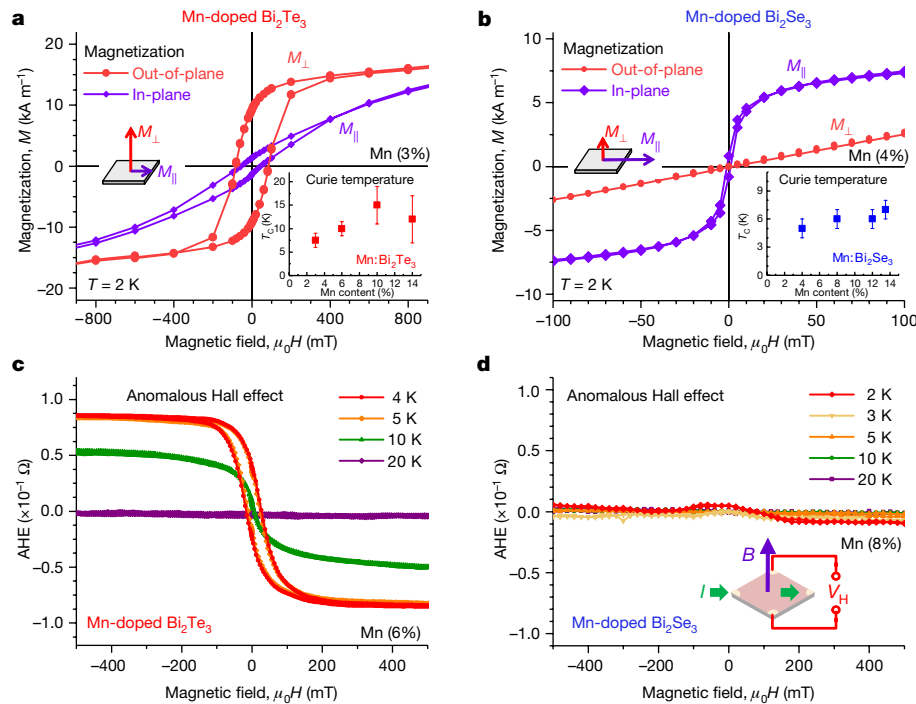


Fig. 2 | Magnetic properties. **a, b**, In-plane and out-of-plane magnetization $M(H)$ of Mn-doped Bi_2Te_3 (**a**) and Bi_2Se_3 (**b**) films with Mn concentrations of 3% and 4% at 2 K, measured with the magnetic field parallel, respectively, perpendicular to the surface, showing a perpendicular anisotropy (easy axis) for Bi_2Te_3 and an in-plane easy axis for Bi_2Se_3 . The insets show the Curie

temperature as a function of Mn concentration. **c, d**, Anomalous Hall effect (AHE) of Mn-doped Bi_2Te_3 (**c**) and Bi_2Se_3 (**d**) measured between 2 K and 20 K. The contribution of the ordinary Hall effect extracted from the high field data was subtracted (see Methods). Owing to the perpendicular magnetic anisotropy, only Mn-doped Bi_2Te_3 displays a pronounced anomalous Hall effect below T_C .

oriented field (M^*) leads to the opposite spin polarization (Fig. 1h, lower panel). This unambiguously proves that the out-of-plane spin polarization below T_C is due to ferromagnetic ordering of the system. Figure 1j shows that, away from the Dirac point, the characteristic helical in-plane spin texture of the parent Bi_2Te_3 is preserved, such that an overall spin texture as displayed in Fig. 1k is formed.

Our measurement of the gap probes the exchange splitting of p electrons of the host material, which ferromagnetically couple to the localized magnetic moments of the Mn ions⁹. The magnitude of the gap thus depends on the exchange coupling, J , and the magnetization, M , along the surface normal direction²¹. Indeed, the gap size nicely follows the temperature dependence of the perpendicular magnetization, M_\perp (Fig. 1k, blue crosses). This clearly demonstrates the direct correlation between the gap and ferromagnetism in the system. In Fig. 1e–g we show that such temperature dependence is not observed for Bi_2Se_3 with a similar Mn concentration of 6%, where instead a large gap of roughly 200 meV exists at all temperatures from 1 K to 300 K (ref. 14). In particular, the gap size does not increase when cooling down to 1 K, well below the T_C of about 6 K (Fig. 1g, inset). This rules out a substantial contribution of magnetism to the Dirac gap for Mn-doped Bi_2Se_3 , in contrast to Bi_2Te_3 , and serves as a cross-check for the magnetic gap opening.

Figure 2 shows the magnetization of Bi_2Te_3 and Bi_2Se_3 for comparable Mn concentrations. The Mn-doped Bi_2Te_3 film shows an easy-axis magnetization normal to the surface: that is, M_\perp is greater than M_\parallel . This perpendicular anisotropy is robust because it does not depend on the Mn concentration (Extended Data Fig. 3) and also occurs for bulk single crystals²². By contrast, for Mn-doped Bi_2Se_3 , the easy axis is parallel to the surface plane (M_\parallel is less than M_\perp) for all investigated Mn concentrations. The coercive field is substantially larger for Mn-doped Bi_2Te_3 than for Mn-doped Bi_2Se_3 , and the anisotropy field at which the in- and out-of-plane magnetizations are equal is two times higher for Mn-doped Bi_2Te_3 (Extended Data Fig. 3). Finally, the ferromagnetic T_C

of Mn-doped Bi_2Te_3 is considerably higher (7–15 K) than for Mn-doped Bi_2Se_3 (5–7 K) (Fig. 2a, b, insets, and Supplementary Information on SQUID measurements). Altogether this shows that Mn-doped Bi_2Te_3 is the more robust and anisotropic ferromagnet. The opposite anisotropy is also revealed by the magnetotransport measurements shown in Fig. 2c, d, where, with magnetic fields applied perpendicular to the films, only Mn-doped Bi_2Te_3 displays a pronounced anomalous Hall effect below T_C , whereas it is negligible in Mn-doped Bi_2Se_3 (Fig. 2d). This perpendicular anisotropy in Mn-doped Bi_2Te_3 is precisely the precondition for the magnetic bandgap opening and the QAHE, whereas an in-plane magnetization as observed for Mn-doped Bi_2Se_3 merely shifts the Dirac cone in momentum parallel to the surface^{9,23}.

Multiscale structure analysis

To clarify how Mn is actually incorporated into Bi_2Te_3 and Bi_2Se_3 , we carried out a multiscale structure analysis for both systems. Figure 3a shows Mn-doped Bi_2Te_3 in high-resolution scanning transmission electron microscopy (HRSTEM). Strikingly, we observe the emergence of a new structure composed of septuple and quintuple layers, instead of the expected periodic sequence of Te–Bi–Te–Bi–Te quintuple layers. The septuple layers consist of the sequence Te–Bi–Te–Mn–Te–Bi–Te, where the Mn atoms predominantly occupy the centre of the septuple²⁴. This self-organized heterostructure formation does not exist for stoichiometric Bi_2Te_3 and Bi_2Se_3 (Extended Data Fig. 4a) and obviously disagrees with the commonly held notion of substitutional Mn incorporation^{9,12,14,22,25}. Figure 3b and Extended Data Fig. 4b show that the same Mn-induced septuple/quintuple heterostructure formation also occurs in Bi_2Se_3 , in agreement with recent observations^{11,26}. Thus, it is a universal mechanism in both material systems. Moreover, we identify this structure as the explanation for the surprisingly small effect^{14,20,27} of Mn doping on the carrier concentration and Fermi level of the system: Mn in MnBi_2Te_4 is electrically neutral because each septuple

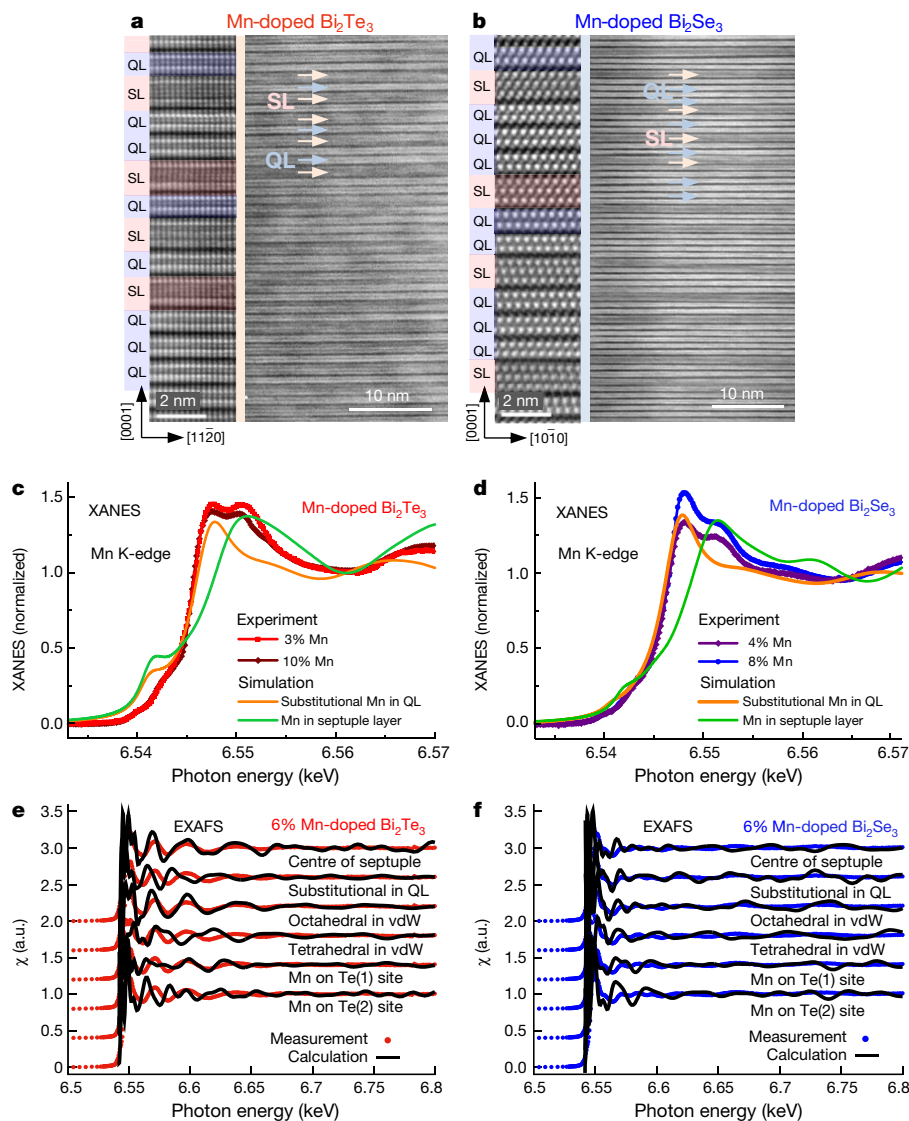


Fig. 3 | Structure analysis by STEM and X-ray absorption spectroscopy.

a, b, STEM cross-sections of Mn-doped Bi_2Te_3 (**a**) and Bi_2Se_3 (**b**), revealing the formation of layered heterostructures consisting of MnBi_2Te_4 (MnBi_2Se_4) septuple layers (SLs) inserted between Bi_2Te_3 (Bi_2Se_3) quintuple layers (QLs). The images were recorded along the $[1100]$ (**a**) and $[1210]$ zone axes (**b**). Owing to the atomic-number contrast, the heavy atoms (Bi) appear brighter in the high-angle annular dark field (HAADF) images. As a result, the septuple layers appear darker in the overview images because of incorporation of the lighter

Mn atoms. The Mn concentration was 10% in **a**, and locally 9% and on average 6% in **b**, according to X-ray diffraction measurements (Fig. 4). **c–f**, Spectroscopic determination of the Mn-incorporation sites in Mn-doped Bi_2Te_3 (**c, e**) and Bi_2Se_3 (**d, f**) by X-ray absorption spectroscopy (XANES in **c, d** and EXAFS in **e, f**) at the Mn K-edge. Experimental data (symbols) are compared with simulations (solid lines) performed for different Mn-incorporation sites in the septuple and quintuple layers, in the van der Waals (vdW) gap or on Te (Se) antisites (see Extended Data Fig. 6).

is formed by insertion of a charge-compensated MnTe double layer into a quintuple layer.

To obtain element-specific information on the Mn-incorporation sites, we carried out X-ray absorption near-edge spectroscopy (XANES) and extended fine structure spectroscopy (EXAFS) at the Mn K-edge (Fig. 3c–f). We analysed the absorption spectra through simulations of all possible Mn-incorporation sites (see Methods section on ‘XANES and EXAFS measurements and simulations’ and Extended Data Figs. 6, 7), showing that Mn in Bi_2Te_3 indeed prefers to be incorporated into the newly formed septuple layers, and that only a minority is incorporated in the quintuple layers. Although the EXAFS data do not completely rule out the incorporation of Mn into octahedral sites in the van der Waals gap, the fact that septuples are never seen in undoped Bi_2Te_3 clearly suggests that the Mn sites are closely linked to the septuple layers. This is supported by high-resolution energy-dispersive X-ray (EDX) elemental maps of the Bi, Te and Mn atoms, in which Mn does

not appear in the van der Waals gaps but mostly in the septuple layers (Extended Data Fig. 4c).

Turning to Mn-doped Bi_2Se_3 we do not observe as intense EXAFS oscillations as for Bi_2Te_3 , indicating that Mn is distributed over different lattice sites, with a larger amount of substitutional Mn and a lesser fraction within the septuple layers. This is highlighted by the XANES spectra, which exhibit a characteristic double-peak structure, with the higher energy peak being attributed to Mn at the centre of the septuple and the lower energy peak to substitutional Mn. Again for Bi_2Se_3 the signal from Mn in the septuples is weaker compared with that in Bi_2Te_3 . For tetrahedrally coordinated interstitial Mn and Mn on Te (Se) antisites, the simulations do not agree with the experiments, indicating that these are not favourable for Mn incorporation. Overall, we conclude that for Bi_2Te_3 the vast majority of Mn is incorporated within the septuple layers and that substitutional Mn is more readily formed in Bi_2Se_3 , especially at lower Mn concentrations.

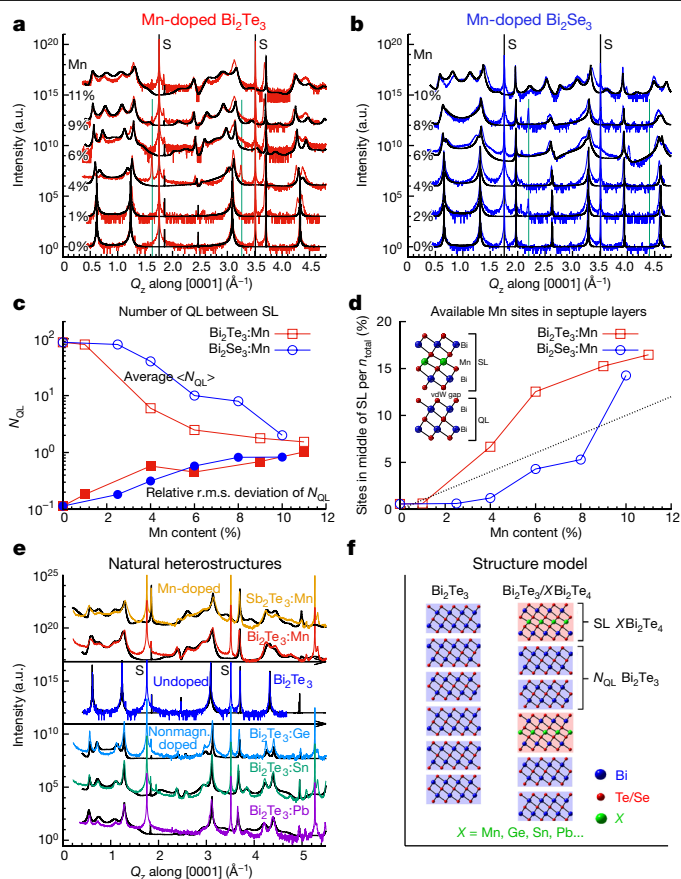


Fig. 4 | X-ray diffraction analysis of self-organized heterostructures in Mn-doped Bi_2Te_3 and Bi_2Se_3 as a function of Mn concentration. The Mn concentration varied from 0% to 11%. **a, b**, Measured diffraction spectra (red and blue lines) recorded along the vertical Q_z direction for Mn-doped Bi_2Te_3 (**a**) and Bi_2Se_3 (**b**) at different Mn concentrations. The diffraction peaks of the substrate are indicated by 'S'. The spectra were fitted using a paracrystal model of statistically varying alternations of Bi_2Y_3 quintuple ($Y = \text{Te}, \text{Se}$) and MnBi_2Y_4 septuple layers (see Methods). An excellent fit (black lines) is obtained for both systems. **c**, The derived average number of quintuples (N_{QL}) between the septuples (open symbols) and the r.m.s. width of the random distribution (filled symbols) are plotted against Mn concentration. A smaller average distance (N_{QL})—that is, a higher concentration of septuples—is found for Bi_2Te_3 than for Bi_2Se_3 . **d**, The number of available Mn sites in the centre of the septuples relative to the total number, n_{tot} , of incorporated Mn atoms is plotted against the nominal Mn concentration. The number expected for unity occupancy is indicated by the dashed line. Experimental points below this line indicate a substantial fraction of Mn atoms residing in other lattice sites. This applies to Bi_2Se_3 but not to Bi_2Te_3 . **e**, Structure analysis for Bi_2Te_3 doped with Ge, Sn or Pb and for Sb_2Te_3 doped with Mn. The good fit of the diffraction data with the paracrystal model (black lines) reveals that the same type of self-organized heterostructures (**f**) is formed in all cases.

To systematically map out the structure evolution on a larger length scale, we carried out X-ray diffraction investigations as summarized in Fig. 4a–d. For both systems we find a pronounced change with increasing Mn concentration, shown by the appearance of additional diffraction peaks that signify the emergence of septuple layers in the structure. The septuples are, however, not incorporated periodically at fixed distances, but rather stochastically after a varying number, N_{QL} , of quintuple layers. This is seen in the STEM cross-sections, where N_{QL} varies between one to seven. To evaluate the diffraction data, we have thus developed an one-dimensional paracrystal model, in which the overall structure is described as a statistically varying sequence of quintuple segments alternating with single septuple layers

(see Methods section ‘X-ray diffraction and simulation with random stacking model’ and Extended Data Fig. 8). Each sequence is characterized by the average number, $\langle N_{\text{QL}} \rangle$, of quintuples between subsequent septuples and the randomness of the statistical N_{QL} distribution—that is, their root mean square (r.m.s.) deviation from the average value.

The model fits (black lines in Fig. 4a, b) show a remarkably good agreement with the diffraction spectra. This corroborates the formation of self-organized quintuple/septuple heterostructures in both systems. From the fits, we obtain the average $\langle N_{\text{QL}} \rangle$ between the septuple layers as well as the r.m.s. deviation as a function of Mn concentration. As shown in Fig. 4c, $\langle N_{\text{QL}} \rangle$ rapidly decreases and thus the density of septuples increases with increasing Mn concentration, underlining that the septuple formation is indeed driven by the Mn doping. This also explains why T_{C} varies so little with the Mn concentration in Fig. 2: the magnetic properties of the samples are largely those of the individual septuple layers. Apparently, in Bi_2Te_3 , the formation of septuple layers starts at a lower Mn concentration than in Bi_2Se_3 and the average separation $\langle N_{\text{QL}} \rangle$ between the septuples is smaller. This difference is highlighted in Fig. 4d, where the number of available Mn sites in the septuples is plotted versus the actual Mn concentration, revealing that in Bi_2Te_3 all Mn atoms can be incorporated in the septuple layers, whereas in Bi_2Se_3 the density of septuples at low Mn concentrations is too small to accommodate all Mn atoms, which must thus be incorporated at other sites as well. We emphasize that our model of self-organized septuple/quintuple heterostructures applies not only to Mn, but also to other nonisovalent dopants such as Ge, Sn and Pb. As a result, very similar diffraction spectra that are well described by the same paracrystal model are obtained, as shown in Fig. 4e, f and Extended Data Fig. 9. This highlights that this new type of incorporation mechanism is completely generic in the tetradymite chalcogenide material systems.

Discussion

The electronic structure of transition-metal impurities in Bi_2Te_3 and Bi_2Se_3 has been studied extensively through density functional theory (DFT) calculations^{9,25}. For Mn in Bi_2Se_3 , a nonmagnetic bandgap of the measured size (200 meV) does not appear in any DFT calculation. Although in principle, depending on orbital symmetry, small gaps of around 4 meV might open even for an in-plane magnetization²⁵, this is obviously much less than what we observe experimentally. The only prediction of a nonmagnetic gap of the magnitude seen in our experiments is from calculations that assume an on-site Coulomb interaction, U , at the impurity site²⁸. On the one hand, Mn forms more substitutional sites in Bi_2Se_3 than in Bi_2Te_3 . They will lead to a larger Coulomb U than for Mn in the centre of the septuple layer, where Mn 3d levels can delocalize in the plane. The size of U , also termed the impurity strength²⁸, indeed affects the nonmagnetic gap: comparing Mn with In doping for Bi_2Se_3 , we find that to reach the same gap size as for 8% Mn, only 2% In is required²⁹. On the other hand, the effect of impurities on the nonmagnetic gap decreases with higher spin–orbit interaction in the host material²⁹, so that Bi_2Te_3 is less susceptible to a nonmagnetic gap opening than Bi_2Se_3 .

To explain the marked difference in the magnetic anisotropy of Mn-doped Bi_2Te_3 and Bi_2Se_3 , we calculated the magnetocrystalline anisotropy for the $\text{Bi}_2\text{Y}_3/\text{MnBi}_2\text{Y}_4$ ($Y = \text{Te}, \text{Se}$) heterostructures (see Methods section ‘DFT calculation of magnetic anisotropy’). In agreement with recent model calculations¹⁰, we find that the strong magnetocrystalline anisotropy favours out-of-plane magnetization in the telluride. In the selenide, however, because of the reduced spin–orbit interaction the magnetocrystalline anisotropy is 3.5 times smaller and practically cancelled by the shape anisotropy. Thus, the higher spin–orbit interaction in the telluride heterostructures turns the magnetization out of the plane and enables the magnetic gap to form at the Dirac point.

Finally, our magnetic gap size of 90 meV for Mn-doped Bi_2Te_3 is five times larger than that predicted theoretically for substitutional Mn⁹. This huge enhancement arises from the naturally formed heterostructure and the enhanced wavefunction overlap of the TSS with the Mn atoms in the MnBi_2Te_4 septuple layer, which supports large magnetic gaps of 38–87 meV, as predicted¹⁰. Mn-doped Sb_2Te_3 displays the same heterostructure formation and out-of-plane magnetic anisotropy as Bi_2Te_3 (Extended Data Fig. 10), and because it is p-type, the Fermi level can be tuned into the magnetic gap by alloying of these systems. This demonstrates the great potential of such structures for stabilizing edge transport in QAHE devices. Theory also suggests that the nontrivial topology is retained in the heterostructures¹⁰, in accordance with the persistence of the Dirac cone surface state and out-of-plane spin texture seen in our ARPES experiments. Therefore, Mn-based topological insulator heterostructures might not only boost edge transport in QAHE devices, but also facilitate the realization of new topological phases such as the axion insulator state^{30,31} and the chiral Majorana fermion³².

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1826-7>.

- Onoda, M. & Nagaosa, N. Quantized anomalous Hall effect in two-dimensional ferromagnets: quantum Hall effect in metals. *Phys. Rev. Lett.* **90**, 206601 (2003).
- Yu, R. et al. Quantized anomalous Hall effect in magnetic topological insulators. *Science* **329**, 61–64 (2010).
- Chang, C.-Z. et al. Experimental observation of the quantum anomalous Hall effect in a magnetic topological insulator. *Science* **340**, 167–170 (2013).
- Checkelsky, J. G. et al. Trajectory of the anomalous Hall effect toward the quantized state in a ferromagnetic topological insulator. *Nat. Phys.* **10**, 731–736 (2014).
- Bestwick, A. J. et al. Precise quantization of the anomalous Hall effect near zero magnetic field. *Phys. Rev. Lett.* **114**, 187201 (2015).
- Chang, C.-Z. et al. High-precision realization of robust quantum anomalous Hall state in a hard ferromagnetic topological insulator. *Nat. Mater.* **14**, 473–477 (2015).
- Grauer, S., Schreyeck, S., Winnerlein, M., Brunner, K., Gould, C. & Molenkamp, L. W. Coincidence of superparamagnetism and perfect quantization in the quantum anomalous Hall state. *Phys. Rev. B* **92**, 201304(R) (2015).
- Tokura, Y., Yasuda, K. & Tsukazaki, A. Magnetic topological insulators. *Nat. Rev. Phys.* **1**, 126–143 (2019).
- Henk, J. et al. Topological character and magnetism of the Dirac state in Mn-doped Bi_2Te_3 . *Phys. Rev. Lett.* **109**, 076801 (2012).
- Otrokov, M. M. et al. Highly-ordered wide bandgap materials for quantized anomalous Hall and magnetoelectric effects. *2D Mater.* **4**, 025082 (2017).
- Hagmann, J. A. et al. Molecular beam growth and structure of self-assembled $\text{Bi}_2\text{Se}_3/\text{MnBi}_2\text{Se}_4$ multilayer heterostructures. *New J. Phys.* **19**, 085002 (2017).
- Xu, S.-Y. et al. Hedgehog spin texture and Berry's phase tuning in a magnetic topological insulator. *Nat. Phys.* **8**, 616–622 (2012).
- Zhang, D. et al. Interplay between ferromagnetism, surface states, and quantum corrections in a magnetically doped topological insulator. *Phys. Rev. B* **86**, 205127 (2012).
- Sánchez-Barriga, J. et al. Nonmagnetic band gap at the Dirac point of the magnetic topological insulator $(\text{Bi}_{1-x}\text{Mn}_x)_2\text{Se}_3$. *Nat. Commun.* **7**, 10559 (2016).
- Scholz, M. R. et al. Tolerance of topological surface states towards magnetic moments: Fe on Bi_2Se_3 . *Phys. Rev. Lett.* **108**, 256810 (2012).
- Ye, M. et al. Quasiparticle interference on the surface of Bi_2Se_3 induced by cobalt adatom in the absence of ferromagnetic ordering. *Phys. Rev. B* **85**, 205317 (2012).
- Sessi, P. et al. Dual nature of magnetic dopants and competing trends in topological insulators. *Nat. Commun.* **7**, 12027 (2016).
- Lee, I. et al. Imaging Dirac-mass disorder from magnetic dopant atoms in the ferromagnetic topological insulator $\text{Cr}_x(\text{Bi}_{0.1}\text{Sb}_{0.9})_{2-x}\text{Te}_3$. *Proc. Natl Acad. Sci. USA* **112**, 1316–1321 (2015).
- Chang, C.-Z. et al. Chemical-potential-dependent gap opening at the Dirac surface states of Bi_2Se_3 induced by aggregated substitutional Cr atoms. *Phys. Rev. Lett.* **112**, 056801 (2014).
- Růžička, J. et al. Structural and electronic properties of manganese-doped Bi_2Te_3 epitaxial layers. *New J. Phys.* **17**, 013028 (2015).
- Rosenberg, G. & Franz, M. Surface magnetic ordering in topological insulators with bulk magnetic dopants. *Phys. Rev. B* **85**, 195119 (2012).
- Hor, Y. S. et al. Development of ferromagnetism in the doped topological insulator $\text{Bi}_{2-x}\text{Mn}_x\text{Te}_3$. *Phys. Rev. B* **81**, 195203 (2010).
- Kharitonov, M. Interaction-enhanced magnetically ordered insulating state at the edge of a two-dimensional topological insulator. *Phys. Rev. B* **86**, 165121 (2012).
- Lee, D. S. et al. Crystal structure, properties and nanostructuring of a new layered chalcogenide semiconductor, Bi_2MnTe_4 . *CrystEngComm* **15**, 5532–5538 (2013).
- Abdalla, L. B., Seixas, L., Schmidt, T. M., Miwa, R. H. & Fazzio, A. Topological insulator Bi_2Se_3 (111) surface doped with transition metals: an ab-initio investigation. *Phys. Rev. B* **88**, 045312 (2013).
- Hirahara, T. et al. Large-gap magnetic topological heterostructure formed by subsurface incorporation of a ferromagnetic layer. *Nano Lett.* **17**, 3493–3500 (2017).
- Lee, J. S. et al. Ferromagnetism and spin-dependent transport in n-type Mn-doped bismuth telluride thin films. *Phys. Rev. B* **89**, 174425 (2014).
- Black-Schaffer, A. M. & Balatsky, A. V. Strong potential impurities on the surface of a topological insulator. *Phys. Rev. B* **85**, 121103(R) (2012).
- Sánchez-Barriga, J. et al. Anomalous behavior of the electronic structure of $(\text{Bi}_{1-x}\text{In}_x)_2\text{Se}_3$ across the quantum phase transition from topological to trivial insulator. *Phys. Rev. B* **98**, 235110 (2018).
- Mogi, M., Kawamura, M., Tsukazaki, A., Yoshimi, R., Takahashi, K. S., Kawasaki, M. & Tokura, Y. Tailoring tricolor structure of magnetic topological insulator for robust axion insulator. *Sci. Adv.* **10**, eaao1669 (2017).
- Xiao, D. Realization of the axion insulator state in quantum anomalous Hall sandwich heterostructures. *Phys. Rev. Lett.* **120**, 056801 (2018).
- He, Q. L., et al. Chiral Majorana fermion modes in a quantum anomalous Hall insulator-superconductor structure. *Science* **357**, 294–299 (2017).

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

Sample growth

Mn-doped Bi_2Te_3 , Bi_2Se_3 , and Sb_2Te_3 layers were grown by molecular beam epitaxy (MBE) on $\text{BaF}_2(111)$ substrates using a Riber 1000 and a Varian GEN II MBE system. Compound Bi_2Te_3 , Bi_2Se_3 , and Sb_2Te_3 as well as elemental sources for Mn, Te and Se were used for control of stoichiometry and composition. For Pb-, Sn- and Ge-doped Bi_2Te_3 , we used additional PbTe, SnTe and GeTe sources. Deposition was carried out at a growth temperature of 330 °C for Bi_2Te_3 and Sb_2Te_3 and 360 °C for Bi_2Se_3 to obtain perfect two-dimensional growth independently of dopant concentrations. This was verified by in situ reflection high-energy electron diffraction (RHEED; Extended Data Fig. 9). Details of growth procedures have been reported previously^{14,20}. Our notation of $x\%$ Mn in Bi_2Te_3 refers to a nominal composition of $(\text{Bi}_{1-x}\text{Mn}_x)_2\text{Te}_{3+3x}$. All layers exhibited n-type conduction with electron concentrations of the order of a few times 10^{19} cm^{-3} (see Extended Data Fig. 5c), except for Mn-doped Sb_2Te_3 which is p-type with a hole concentration of a few times 10^{20} cm^{-3} . Immediately after growth, samples used for ARPES were capped in situ with amorphous Se and Te capping layers at room temperature to protect the surface against oxidation. This cap was removed just before the ARPES experiments by in situ sputtering and annealing.

ARPES

Photoemission experiments were performed with the ARPES-1³ end station at the UE112-PGM2b undulator beam line of the BESSY II synchrotron radiation source. The lowest reachable temperature is 1 K. The experimental geometry has the following characteristics: with the central axis of the analyser lens and the polar rotation axis of the sample defined as the x and z axes of a spherical coordinate system, the photons impinge the sample under an azimuthal angle φ of 45° and a polar angle of 84°. The light polarization is horizontal (along the x axis). The entrance slit of the hemispherical analyser is placed parallel to the z axis. The measurements at $h\nu = 50 \text{ eV}$ were performed with an energy resolution of 10 meV. The temperature-dependent leading-edge shifts in Fig. 1c and Extended Data Fig. 1a are obtained by approximating the photoemission intensity in the indicated ranges with a line. The slope of this line is constrained to be identical for the low- and high-temperature spectra within the same section. Our notation Δ refers to the full gap.

The measurement of the magnetic exchange splitting at the Dirac point through the temperature dependence in ARPES is in a certain sense analogous to the case of gadolinium metal. In both cases the magnetic coupling of localized magnetic moments ($3d$ in Mn, $4f$ in Gd) is mediated by itinerant electrons ($5p$ in Te, $5d$ in Gd) which can be probed by ARPES. At the Dirac point, electronic states of the Te $5p$ character are probed. In Gd, the $5d$ band probed by ARPES splits by 0.85 eV when the temperature is lowered from $1.02 T_c$ to $0.27 T_c$ ($T_c = 293 \text{ K}$ for bulk Gd)³³.

Spin-resolved ARPES

Spin-resolved ARPES was measured at the RGL2 end station at the U125/2 undulator beamline of BESSY II. It comprises a Scienta R4000 hemispherical analyser with two Mott-type spin polarimeters operated at 26 kV (ref. ³⁴). The lowest temperature is 6.5 K. Light is incident under an azimuthal angle of 45° and a polar angle of 90° (Extended Data Fig. 2a). The light polarization is horizontal (along the x axis). The spin polarimeter detects the out-of-plane and one in-plane component of the spin polarization. The measured in-plane projection is tangential to the Dirac cone and perpendicular to the analyser entrance slit and electron momentum. The out-of-plane component lies within the electron emission plane and is parallel to the sample normal along the z -direction. The angular resolution was 0.75° and the energy resolution of the measurement at a photon energy of 30 eV was set to 45 meV. We note that the measurement of the spin splitting is not limited by the energy resolution because the spin-up and spin-down channels

count independently of each other. An example of this is the exchange splitting of a Ni(111) surface state, measured by spin-resolved inverse photoemission as $18 \pm 3 \text{ meV}$ at an energy resolution of 300–400 meV (ref. ³⁵).

Transport measurements

Temperature-dependent transport measurements were performed in van der Pauw geometry with out-of-plane magnetic fields ranging from 0 T to 2 T and temperatures down to 1 K using a Cryogenic mini cryogen-free system. Extended Data Fig. 5 shows the complete data set (Hall resistance plotted against magnetic field, as well as carrier concentration, n , and carrier mobility, μ , plotted against temperature) for the Mn-doped Bi_2Te_3 and Bi_2Se_3 films with respectively 6% and 8% Mn. Below the ferromagnetic transition T_c (10 K and 6 K, respectively), the Hall resistance comprises the contribution from the ordinary Hall effect (proportional to $1/ne$) and the anomalous Hall effect (proportional to the magnetization, M). Because the latter is proportional to the perpendicular magnetization, the anomalous Hall contribution is minute for Mn-doped Bi_2Se_3 films, for which the magnetization vector is nearly parallel to the film plane. Above T_c the anomalous Hall contribution is absent.

Magnetic characterization

The magnetic properties were determined by measuring magnetization, M , as a function of the applied external field, H , and as a function of temperature, T (ranging from 2 K to 300 K), using a superconducting quantum interference device (SQUID) magnetometer (Quantum Design MPMS-XL5). We determined the Curie temperature from the $M(T)$ curves as exemplified in Extended Data Fig. 3. Note that we do not observe an enhancement of T_c at the surface probed by XMCD¹⁴. The magnetic field was applied either parallel (out-of-plane) or perpendicular (in-plane) to the c axis of the films. The diamagnetic contribution of the $\text{BaF}_2(111)$ substrate was determined from the slope of the $M(H)$ curve recorded at 300 K in high magnetic fields, and was subtracted from the raw data. Identical sample pieces were used for in-plane and out-of-plane measurements. The sample size was $4 \times 4 \text{ mm}^2$.

Scanning transmission electron microscopy

Atomic resolution HRSTEM images were obtained with a FEI Titan G2 60-300 STEM equipped with a Cs probe corrector and a FEI Titan 60-300 Themis equipped with a Cs image corrector, which were operated at 300 keV. The HRSTEM data were recorded with a HAADF detector and the images processed using a Wiener filter for noise minimization³⁶. Thin cross-sectional lamellae from Mn-doped Bi_2Te_3 and Bi_2Se_3 films with Mn concentrations of 10% and 6%, respectively, were prepared by focused ion beam (FIB) milling (ZEISS Crossbeam XB 1540 and FEI Helios NanoLab 660) along two different crystallographic directions of the $\text{BaF}_2(111)$ substrate ($[\bar{2}11]$ and $[0\bar{1}1]$, respectively). Owing to the epitaxial relationship between the films and the $\text{BaF}_2(111)$ substrate, this yields $[\bar{1}100]$ and $[12\bar{1}0]$ zone axes with respect to the Bi_2Te_3 and Bi_2Se_3 layers. Pre-characterization of the lamellae and overview STEM images were obtained with a JEOL JEM-2200FS STEM operated at 200 keV. Shortly before the HRSTEM images were taken, the lamellae were additionally thinned to remove the amorphous surface layers and damaged regions caused by the initial FIB preparation using a Fischione 1040 NanoMill. To map the element distribution (Extended Data Fig. 4c), we carried out energy-dispersive X-ray analysis using a Bruker Super-X detector.

XANES and EXAFS measurements and simulations

The XANES and EXAFS spectra at the Mn K-edge were recorded at, respectively, the ID12 and BM23 beamlines of the European Synchrotron Radiation Facility in total fluorescence yield³⁷. The isotropic XANES spectrum was derived from a weighted average of two XANES spectra recorded with two orthogonal linear polarizations parallel

and perpendicular to the *c* axis of the film. The resulting X-ray linear dichroism spectra corroborate the findings of XANES and EXAFS with regard to Mn incorporation.

For XANES simulations, we used the FDMNES code³⁸ with a multiple scattering approach on a muffin-tin potential, for a supercell comprising the nominal bulk Bi₂Te₃ or Bi₂Se₃ lattices with a Mn atom replacing one Bi atom within the quintuple layers (substitutional Mn), and with Mn incorporated in the central layers of the septuples (Extended Data Fig. 6). Note that placing Mn as an octahedral interstitial within the van der Waals (vdW) gap leads to somewhat similar results as with Mn in the central position of the septuple layer, while tetrahedral Mn interstitials were in lesser agreement with experiment (Fig. 3c–f). Concerning the pre-edge features in the XANES data, it is known that the FDMNES code using the multiple scattering formalism has difficulties in reproducing the $3d-4p$ hybridized states at the pre-edge feature well; nevertheless the main absorption features are well reproduced and we draw conclusions only from that spectral region.

We used identical input geometries for the EXAFS and XANES simulations. We measured EXAFS spectra for a series of Mn-doped Bi₂Te₃ and Bi₂Se₃ samples with Mn concentrations ranging from 4% to 13%. We fitted the EXAFS data at the Mn K-edge using the FEFF9 code, assuming Mn at different lattice sites. The spectra with a model of Mn atoms in the centre position of the septuple layer are shown in Extended Data Fig. 7. The first coordination shell includes six anion atoms in the octahedral environment. We note again that the substitutional position and the interstitial position in the van der Waals gap have octahedral coordination. The distances of the nearest neighbours in the first coordination shell derived from the fits are listed in Extended Data Fig. 7e. We note that Mn atoms substituting Bi, Te(1) or Se(1) atoms in the quintuple layers (using the notation of Extended Data Fig. 6) have two different neighbours with different distances.

X-ray diffraction and simulation with random stacking model

We determined the crystal structure using symmetric X-ray diffraction scans and reciprocal space maps in the vicinity of the (10 $\bar{1}$.20) reciprocal lattice point. The measurements were performed using a Rigaku SmartLab diffractometer with a copper X-ray tube and channel-cut Ge(220) monochromator. Symmetric scans along the [000.1] reciprocal space direction (*c* axis) were fitted with a modified one-dimensional paracrystal model³⁹, in which random sequences of Bi₂Te₃ (or Sb₂Te₃ or Bi₂Se₃) quintuple segments alternate with MnBi₂Te₄ (or MnSb₂Te₄ or MnBi₂Se₄) septuples along the *c* axis. For the samples doped with *X* = Pb, Sn or Ge (Fig. 4e), the septuples consist of *X*Bi₂Te₄. For the quintuples, the spacings of atomic planes were set to the nominal values of Bi₂Te₃ and Bi₂Se₃, and for the septuples the distance of the Mn (or Pb, Sn or Ge) plane to the nearest neighbour Te (or Se) was set to correspond to the nearest-neighbour distances determined by EXAFS.

The random sequences of quintuple and septuple layers are generated using the following assumptions. First, the length of the individual quintuple segments, N_{QL} , is given by the gamma distribution with a certain mean value, $\langle N_{\text{QL}} \rangle$, and a r.m.s. deviation, σ (r.m.s.d.); we show its relative value, $\sigma/\langle N_{\text{QL}} \rangle$. The septuple segments always consist of only one single septuple layer—that is, septuple layers are not positioned next to each other. Note that we carried out additional test fits assuming a variable length of septuple segments, but the best fits tend to the result with just one septuple layer embedded in the blocks of quintuple layers. Thus, we fixed the length of the septuple segments to one, in order to keep the number of fitting parameters as small as possible.

Second, we set the distances of the individual atomic planes in the quintuples to the values of the pure Bi₂Y₃ phases (*Y* = Te, Se), as we described previously³⁹. For the septuples we have set the distances of the next Te or Se anion sites to the Mn in the central planes to correspond to the distances determined by EXAFS. The total thickness of the septuple equals approximately $4/3$ of d_{QL} ; that is, $d_{\text{SL}} = 4/3d_{\text{QL}}$.

This relation is almost exactly satisfied for Bi₂Te₃, whereas for Bi₂Se₃ $d_{\text{SL}} = 1.02 \times 4/3d_{\text{QL}}$.

For the random sequences generated in this way, we calculated the XRD diffraction spectra using the one-dimensional paracrystal model described before³⁹, and compared the spectra to the experimental data. Examples of simulated profiles with various parameter values are shown in Extended Data Fig. 8a–d. Extended Data Fig. 8a, b depict the influence of the average number of quintuples, $\langle N_{\text{QL}} \rangle$, between the septuples for a fixed disorder, that is, $\sigma/\langle N_{\text{QL}} \rangle = 0.5$. High values of $\langle N_{\text{QL}} \rangle$ correspond to an almost pure Bi₂Y₃ lattice with just few septuples present in the stack. Such a system corresponds to samples with low Mn doping. Smaller values of $\langle N_{\text{QL}} \rangle$ lead to a multilayer system, in which additional satellite diffraction peaks appear. The limiting case of $\langle N_{\text{QL}} \rangle = 1$, which is on average one quintuple alternating with one septuple, corresponds to the top blue line, with the peak positions corresponding to an average periodicity $P = d_{\text{QL}} + d_{\text{SL}}$ along the growth direction. Extended Data Fig. 8c, d show the influence of the randomness (r.m.s.) on the diffraction spectra for a constant $\langle N_{\text{QL}} \rangle$ of 5. A small r.m.s. corresponds to a periodic multilayer of quintuple and septuple segments, with corresponding sharp superlattice maxima, while large r.m.s. values correspond to a disordered system with accordingly smeared profiles. For the samples doped with *X* = Pb, Sn or Ge (Fig. 4e), the fitted paracrystal parameters are listed in Extended Data Fig. 9g.

We note that the relation $d_{\text{SL}} = 4/3d_{\text{QL}}$ has quite an important consequence for the diffraction spectra, because there is a single periodicity that is common to both quintuples and septuples. The simulated and experimental diffraction profiles shown in Fig. 4 have sharp peaks corresponding to such a periodicity independently of the statistical ordering of quintuple and septuple segments. The corresponding peaks appear at the positions (000.9) and (000.18) of the Bi₂Y₃ structure. The average interplanar distance in the septuples is smaller than in the quintuples, because the septuple has $7/5 = 1.4$ more atomic planes but is only thicker by a factor of $4/3 = 1.33$.

We have also determined the Mn-concentration dependence of the in-plane lattice parameter *a* of the Mn-doped Bi₂Y₃ layers from asymmetric reciprocal space maps recorded in the vicinity of the (101.20) reciprocal lattice point of the Bi₂Y₃ structure. These results, as well as those for the average interplanar distance in the *c* axis direction, $\langle d \rangle$, are plotted in Extended Data Fig. 8e, f. In Mn-doped Bi₂Te₃, we observe a decrease in both lattice parameters with increasing Mn content. This can be explained by the fact that a higher concentration of septuple layers leads to a smaller average interplanar distance, while for Mn-doped Bi₂Se₃ the Mn content has less influence on $\langle d \rangle$ owing to the smaller number of septuple layers formed. In fact, in Bi₂Se₃ we do not observe any concentration dependence of the interplanar distance $\langle d \rangle$ up to Mn concentrations of 8%. This is in agreement with the finding from X-ray diffraction that, for low Mn contents in Bi₂Se₃, only a very low number of septuples (Fig. 4) is present. Mn atoms also cause a small shrinking of the in-plane lattice parameter *a* for both Bi₂Te₃ and Bi₂Se₃ (Extended Data Fig. 8e, f).

DFT calculation of magnetic anisotropy

To get a reliable value for the magnetic anisotropy of Mn-doped Bi₂Y₃ (*Y* = Se, Te) with septuple/quintuple layer structure, we carry out ab initio calculations using the well established full-potential linearized augmented plane wave (FLAPW) method, as implemented in the WIEN2k code⁴⁰. Our calculations are based on the local density approximation. Here we used experimental lattice parameters for the multilayer system with alternating septuple/quintuple layers, which corresponds to an Mn concentration of 8%. One part of the magnetic anisotropy energy, known as magnetocrystalline anisotropy (MCA), arises from spin–orbit coupling; the other part, the so-called shape anisotropy, E_{shape} , comes from the magnetic dipole–dipole interaction of the individual magnetic moments. It is well known that a sufficiently dense mesh in the Brillouin zone is important for *k*-space integration.

On the basis of this insight, we used a $45 \times 45 \times 7$ Monkhorst–Pack grid in the full Brillouin zone⁴¹. In our benchmark calculation for FePt⁴² we showed that, in addition to the convergence of k -points, it is important to incorporate all FLAPW eigenfunctions when spin–orbit coupling is included as an additional term to the scalar-relativistic Hamiltonian, the so-called second variational step⁴³. This basis set is controlled by energy parameters (E_{\min} and E_{\max}). To achieve a high accuracy, we set E_{\min} to be -10 Ry and E_{\max} to be 5 Ry. To obtain a stable value for the MCA, the energy parameters, E_{ϕ} , used for calculating radial wavefunctions, $u_{\phi}(r, E_{\phi})$, are determined very precisely—that is, to better than 0.1 mRy.

As the WIEN2k code solves the Dirac equation in an approximate way, it is necessary to check the reliability of these sensitive calculations. For cross-checking we used the multiple-scattering KKR Green function method as implemented in the SPRKKR code^{44,45}. This scheme solves the proper Dirac equation; hence the relativistic effects are fully included in SPRKKR, unlike with the second variational step.

Obtaining the MCA energy by subtracting the total energies is computationally very costly. The need for self-consistent calculations for two magnetization directions can be avoided if one relies on the magnetic force theorem. In this approach the MCA energy is calculated using a frozen spin-dependent potential⁴⁶. The MCA energy is then obtained by subtracting the band energies. E_{shape} is calculated using classical electromagnetic theory. Good agreement has been found between WIEN2k and SPRKKR codes for the MCA energy calculation. For the sake of simplicity we present only the WIEN2k results (Extended Data Fig. 3g).

Ge-, Sn- and Pb-doped Bi₂Te₃ and Mn-doped Sb₂Te₃

Natural formation of quintuple/septuple layer heterostructures is a generic and universal feature in Bi- and Sb-based tetradymite chalcogenide topological insulators doped with elements that prefer a 2^+ state. This is demonstrated by a series of complementary thin film samples consisting of Bi₂Te₃ doped with Ge, Sn, and Pb, as well as of Sb₂Te₃ doped with Mn, using similar growth conditions to those described above. In all cases, two-dimensional growth was observed under the given growth conditions, as shown by the RHEED patterns recorded *in situ* during growth and depicted in Extended Data Fig. 9a–f. The structure of the films was analysed by X-ray diffraction, and the measured diffraction spectra were fitted with the same random stacking paracrystal model described in the section ‘X-ray diffraction and simulation with random stacking’, assuming that the additional doping element ($X = \text{Pb, Ge, Sn, Mn}$) induces the same type of septuple layer formation in which an additional $X\text{Te}$ double layer is inserted. As shown in Fig. 4e, in all cases an excellent fit of the diffraction spectra is obtained with our paracrystal structure model. Moreover, for the same concentration of dopant elements, the average number of quintuple layers, $\langle N_{\text{QL}} \rangle$, between these additionally inserted septuple layers and its r.m.s. variation, σ_{QL} , derived from the model fits (see Extended Data Fig. 9g) turn out to be nearly the same for each dopant element.

Data availability

The data sets generated and analysed here are available from the corresponding authors on reasonable request.

Code availability

The code for the paracrystal model is available from the corresponding authors upon request. The electronic structure codes Wien2K and

SPR-KKR and X-ray absorption fine structure codes FDMNES and FEFF9 can be downloaded after the corresponding licence requirements given on the respective webpages are fulfilled.

33. Kim, B., Andrews, A. B., Erskine, J. L., Kim, K. J. & Harmon, B. N. Temperature dependent conduction-band exchange splitting in ferromagnetic hcp gadolinium: theoretical predictions and photoemission experiments. *Phys. Rev. Lett.* **68**, 1931–1934 (1992).
34. Burnett, G. C., Monroe, T. J. & Dunning, F. B. High-efficiency retarding-potential Mott polarization analyzer. *Rev. Sci. Instrum.* **65**, 1893–1896 (1994).
35. Passek, F. & Donath, M. Spin-split image-potential-induced surface state on Ni(111). *Phys. Rev. Lett.* **69**, 1101–1104 (1992).
36. Mitchell, D. HRTEM Filter. *Austrian Centre for Electron Microscopy and Nanoanalysis* https://dm-script.tugraz.at/dm/source_codes/181 (2007).
37. Rogalev, A., Wilhelm, F., Goulon, J. & Goulon-Ginet, C. C. in *Magnetism and Synchrotron Radiation: Towards the Fourth Generation Light Sources* vol. 151 (eds Beaupaire, E., Bulou, H., Joly, L. & Scheurer, F.) 289–314 (Springer, 2013).
38. Bunău, O. & Joly, Y. Self-consistent aspects of x-ray absorption calculations. *J. Phys. Condens. Matter* **21**, 345501 (2009).
39. Steiner, H. et al. Structure and composition of bismuth telluride topological insulators grown by molecular beam epitaxy. *J. Appl. Cryst.* **47**, 1889–1900 (2014).
40. Blaha, P., Schwarz, K., Madsen, G. K. H., Kvasnicka, D. & Luitz, J. *Wien2k, an augmented plane wave plus local orbital program for calculating the crystal properties*. <http://www.wien2k.at> (2001).
41. Monkhorst, H. J. & Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **13**, 5188–5192 (1976).
42. Khan, S. A., Blaha, P., Ebert, H., Minár, J. & Sipr, O. Magnetocrystalline anisotropy of FePt: a detailed view. *Phys. Rev. B* **94**, 144436 (2016).
43. MacDonald, A. H. & Vosko, S. H. A relativistic density functional formalism. *J. Phys. C Solid State Phys.* **12**, 2977–2990 (1979).
44. Ebert, H., Ködderitzsch, D. & Minár, J. Calculating condensed matter properties using the KKR-Green’s function method—recent developments and applications. *Rep. Prog. Phys.* **74**, 096501 (2011).
45. Ebert, H. The Munich SPRKKR package, version 7. <http://olymp.cup.uni-muenchen.de> (2012).
46. Mackintosh, R. & Andersen, O. K. in *Electrons at the Fermi surface* Vol. 3 (ed. Springford, M.) 149–222 (Cambridge Univ. Press, 1980).

Acknowledgements ARPES experiments were performed at BESSY II of Helmholtz-Zentrum Berlin and the EXAFS and XANES experiments at the European Synchrotron Radiation Facility. We thank B. Henne, F. Wilhelm and A. Rogalev for support with XANES and EXAFS measurements; W. Grafeneder and G. Hesser for TEM sample preparation; V. Holý for advice on the paracrystal model; and G. Bihlmayer and A. Ernst for helpful discussions. This project was supported by the Austrian Science Fund (FWF, project P30969-N27 and P28185-N27); the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development in the frame of the Christian Doppler Laboratory for Nanoscale Phase Transformations; the Deutsche Forschungsgemeinschaft (grants SPP 1666, SFB 1143 project C4, SFB 1277 project A2); the Central European Institute of Technology (CEITEC) Nano research infrastructure (ID LM2015041, MEYS CR, 2016–2019) and Computational and Experimental Design of Advanced Materials with New Functionalities (CEDAMNF; grant CZ.02.1.01/0.0/0.0/15_003/0000358) of the Czech Ministerstvo Školství Mládeže a Tělovýchovy (MSMT); the Impuls- und Vernetzungsfonds der Helmholtz-Gemeinschaft (Virtual Institute New States of Matter and their Excitations and Helmholtz-Russia Joint Research Group no. HRSF-0067); and the European Union Horizon 2020 programme (grant 823717-ESTEEM3).

Author contributions Samples were grown by S.W., H.S., V.V.V. and G.S. X-ray analysis was carried out by S.W., H.S., G.S., J.R. and O.C. O.C. performed paracrystal modelling and magnetotransport measurements. XANES and EXAFS measurements were made by A.N., O.C., H.S. and G.S., and the simulations by O.C., A.N., J.R. and J. Minár. SQUID was carried out by A.N., and HR-STEM by M.A., H.G., S.W., G.K., O.C. and J. Michalička. DFT calculations were done by S.A.K., J. Minár and H.E. ARPES was carried out by E.D.L.R. and P.S.M., and spin-resolved ARPES by J.S.-B., F.F. and A.V. The work was coordinated by G.S., G.B. and O.R. The manuscript was written by O.R. and G.S. with input from all authors.

Competing interests The authors declare no competing interests.

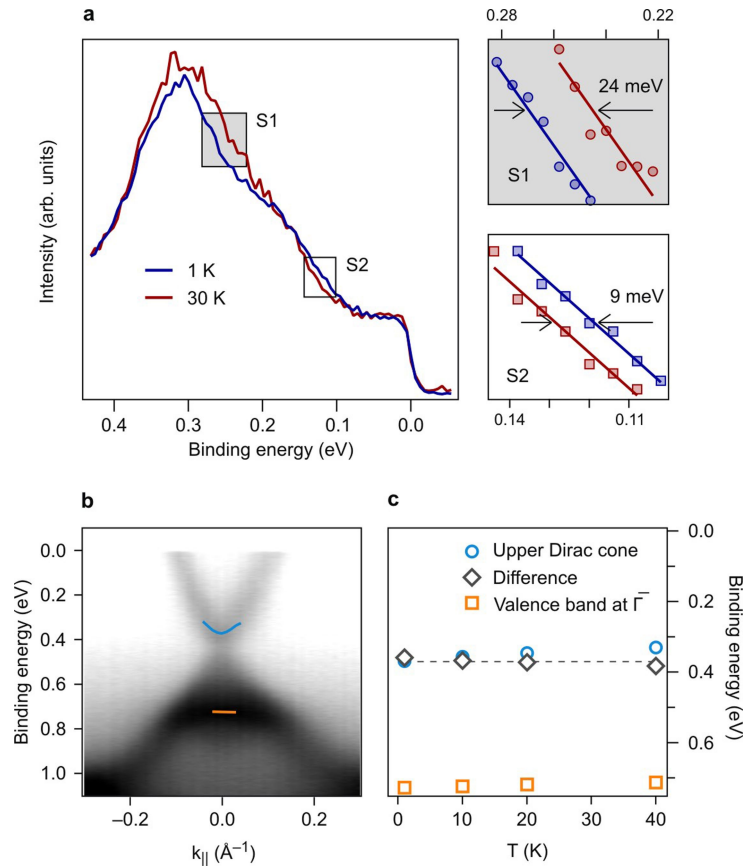
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1826-7>.

Correspondence and requests for materials should be addressed to O.R. or G.S.

Peer review information *Nature* thanks Alexander Balatsky, Jacek Furdyna and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

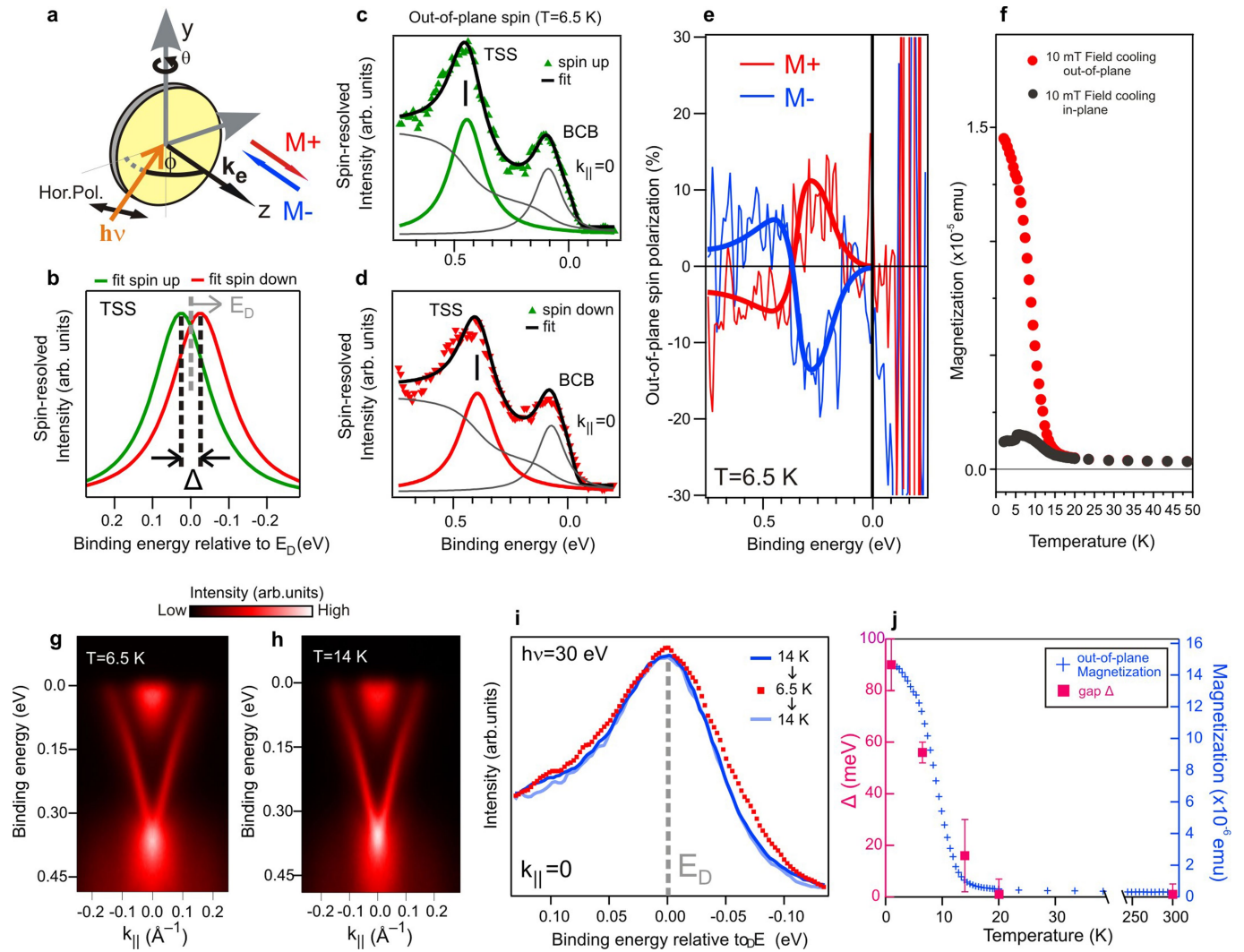
Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | ARPES measurements. **a**, Data for an additional Mn-doped Bi_2Te_3 film with Mn concentration of 10%. The normal emission spectra shown on the left, recorded at 1 K and 30 K, show a substantial redistribution of spectral weight around the binding energy of approximately 180 meV when crossing the ferromagnetic transition at $T_c = 12$ K. The shifts in the regions marked S1 and S2, shown on the right on a magnified scale, are of similar magnitude to that seen for the 6% Mn-doped case in Fig. 1. The shifts marked by arrows are compatible with a 100 meV gap opening at the Dirac point. ARPES was measured with p-polarized light and $h\nu = 50$ eV. **b**, **c**, ARPES measurements, showing that the gap in 6% Mn-doped Bi_2Se_3 is independent of temperature. **b**, ARPES $E(k)$ map recorded at 1 K, with the angle-dependent binding energies

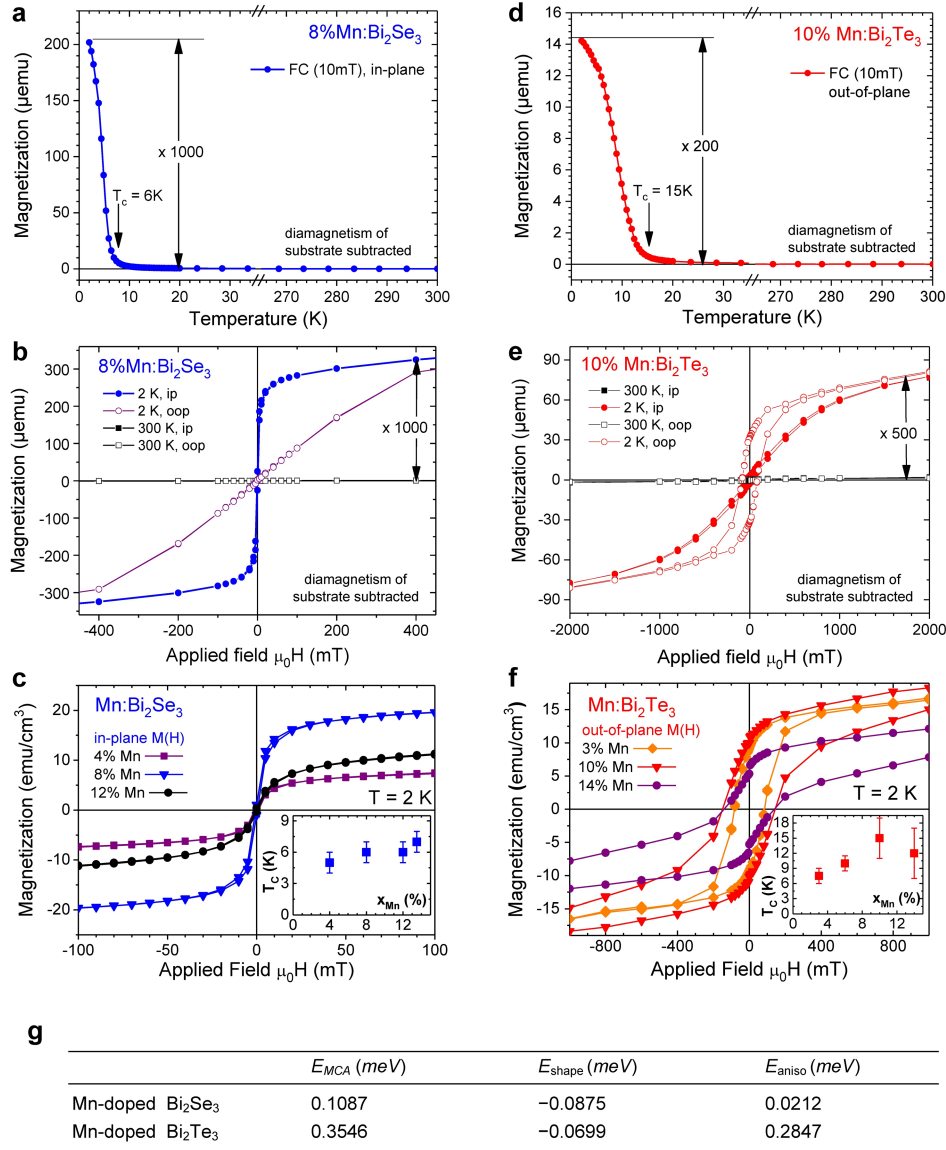
of the upper Dirac cone and bulk valence band indicated with blue and orange lines, obtained from Lorentzian fits to energy-distribution curves.

c, Temperature dependence of the binding energies of the upper Dirac cone minimum (blue circles), the bulk valence band at Γ^- (orange squares) and their difference (black diamonds). The ferromagnetic Curie temperature is 6 K as obtained by SQUID. This analysis represents an alternative to that in Fig. 1g, which was based on fits of the upper and lower Dirac cones. In both cases, the data do not provide any indication of a relative or absolute shift of the band edges, or of a gap of the order seen in Mn-doped Bi_2Te_3 when crossing the ferromagnetic transition temperature.



Extended Data Fig. 2 | Spin-resolved ARPES of Mn-doped Bi₂Te₃. **a**, Geometry of the spin-resolved ARPES experiments, including the magnetization directions indicated by M^+ and M^- . Hor. Pol., horizontal light polarization. **b**, Plot of the fit results from spin-up and spin-down spectra of the topological surface state (TSS) at the Dirac point (E_D), and determination of the magnetic exchange splitting, $\Delta = 56 \pm 4$ meV, at 6.5 K. **c**, **d**, Fit to the spin-resolved spectra at 6.5 K, including a transition from the bulk conduction band (BCB). **e**, Demonstration that the spin polarization reverses when the magnetization, M , is reversed. This reversal was achieved by field cooling in an applied field of 10 mT. **f**, Temperature-dependent magnetization, $M(T)$, measured by SQUID on a reference sample that was identical to that used to determine the magnetic field necessary for field cooling and magnetization reversal. **g**, **h**, Before the spin-resolved ARPES measurement, the reversible temperature-dependent

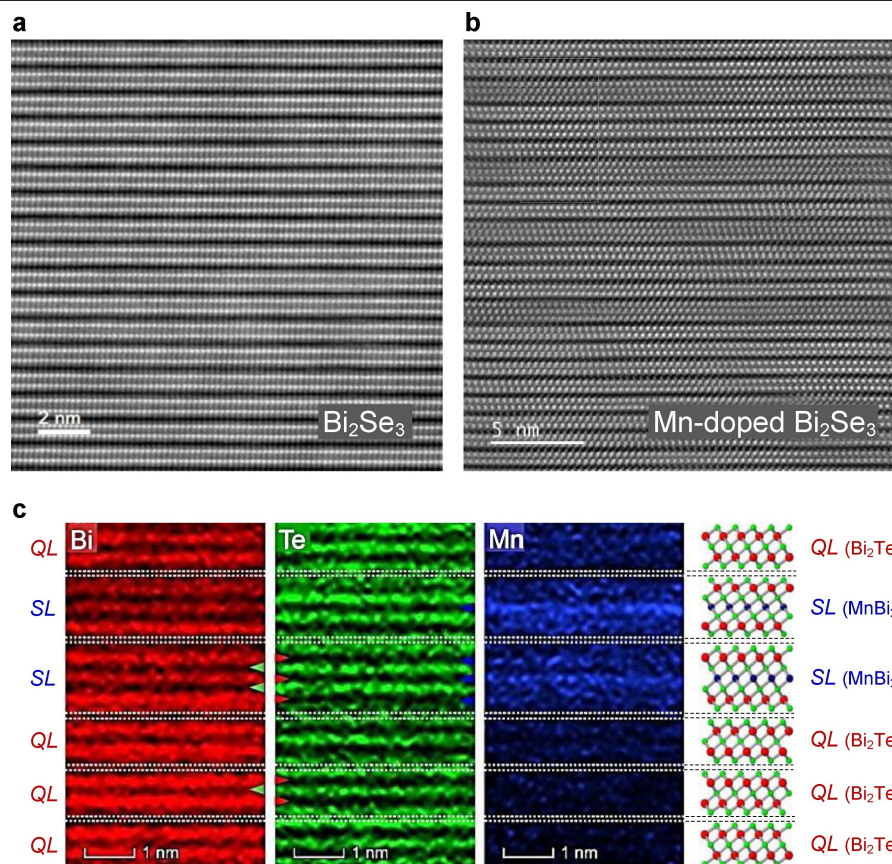
broadening in ARPES was verified by cooling from 14 K to 6.5 K and warming up again to 14 K. At a photon energy of 30 eV, most of the intensity near the Fermi energy stems from the bulk conduction band. **i**, Reversible broadening of the energy-dispersion curves upon cooling from above to below T_c and warming up again. **j**, TSS gap at the Dirac point Δ derived by ARPES and spin-resolved ARPES (red squares), plotted against temperature, together with the temperature-dependent out-of-plane magnetization (blue crosses), showing that the magnetic exchange splitting at the Dirac point Δ faithfully follows the magnetization perpendicular to the sample surface. Data at 1 K and 20 K are from Fig. 1c, d. Data from spin-resolved photoemission (at 6.5 K and 300 K) have the smallest error. Data for 14 K were derived from **i**, taking the spin-resolved data from 6.5 K as a reference point.



Extended Data Fig. 3 | Magnetic properties of Mn-doped Bi₂Se₃ and Bi₂Te₃.

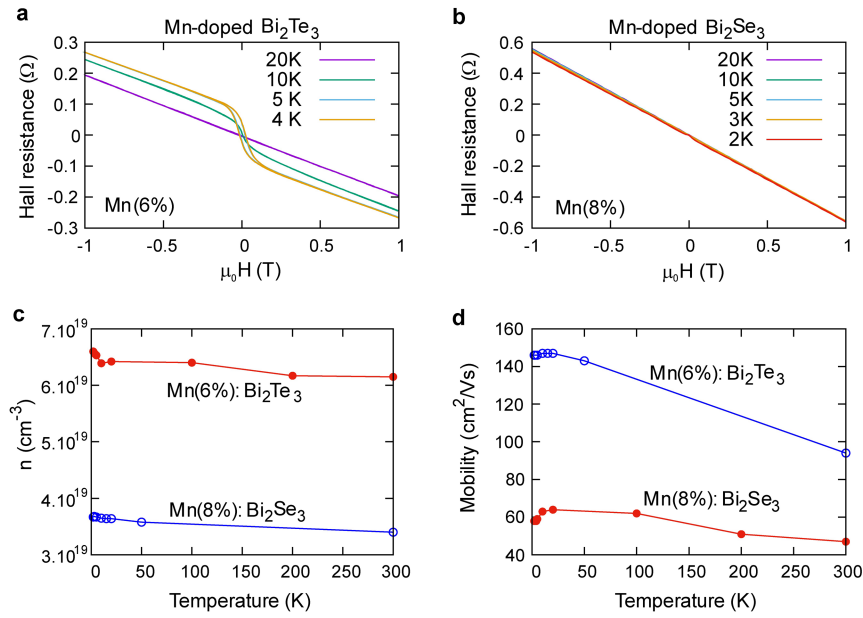
a, d, Temperature-dependent magnetization, $M(T)$, used to determine the ferromagnetic Curie temperature, T_c , Mn-doped Bi₂Se₃ (**a**) and Bi₂Te₃ (**d**). The magnetization was measured after field cooling (FC) at 10 mT. As indicated, below T_c the magnetization of the samples rises steeply by more than two orders of magnitude. **b, e**, In-plane (ip) versus out-of-plane (oop) hysteresis loops at 300 K and 2 K, showing the absence of ferromagnetism at room temperature. **c, f**, Magnetization versus applied field, $M(H)$, for samples with different Mn concentrations, x_{Mn} , as indicated. For all Mn-doped Bi₂Se₃ films,

the easy axis of magnetization is found to be in plane, whereas for all Mn-doped Bi₂Te₃ films it is perpendicular to the surface. The insets show the Curie temperature, T_c , plotted against Mn concentration. For all measurements, the diamagnetic contribution of the substrate measured at 300 K was subtracted. **g**, Magnetocrystalline anisotropy energy, E_{MCA} , obtained through DFT for Bi₂Se₃/MnBi₂Se₄ and Bi₂Te₃/MnBi₂Te₄ by subtracting band energies for two orientations of the magnetization. Shown are magnetocrystalline anisotropy $E_{MCA} = E^{(M||x)} - E^{(M||z)}$, shape anisotropy E_{shape} and total magnetic anisotropy $E_{aniso} = E_{MCA} + E_{shape}$.



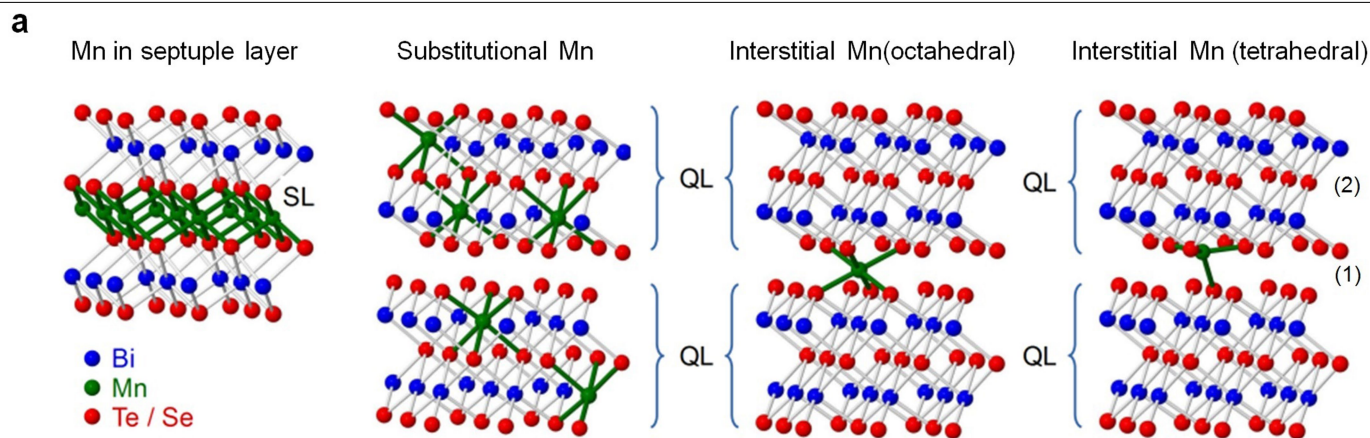
Extended Data Fig. 4 | STEM of pure and Mn-doped Bi_2Se_3 and EDX maps of Mn-doped Bi_2Te_3 . **a, b**, Comparison of HR-STEM HAADF cross-sections of Bi_2Se_3 (**a**) and Mn-doped Bi_2Se_3 ($x_{\text{Mn}} = 6\%$) (**b**) films grown under identical growth conditions, showing the high structural perfection and that additional septuple layers are formed only with Mn doping, whereas the pure Bi_2Se_3 film consists only of quintuple layers. These STEM cross-sections were recorded along two different zone axes. **c**, Atomic-layer-resolved distribution of the Bi,

Te and Mn atoms of a Mn-doped Bi_2Te_3 film ($x_{\text{Mn}} = 10\%$) obtained by STEM-EDX mapping. The Mn atoms are predominantly incorporated in the centre of the septuple layers and to a lesser extent in the outer layers of the septuple units. No Mn is seen in the van der Waals gaps. Note that in this sample, because of the higher Mn concentration, two subsequent septuples are observed in the STEM cross-section.



Extended Data Fig. 5 | Anomalous Hall effect. Data for Mn-doped Bi_2Te_3 and Bi_2Se_3 with respectively 6% and 8% Mn. **a, b**, Raw data for Hall resistance as a function of magnetic field applied perpendicularly to the surface, measured at different temperatures above and below T_c as indicated. **c, d**, Temperature

dependence of the carrier concentration and Hall mobility. Note that above T_c the contribution of the anomalous Hall effect to the Hall voltage is negligible and therefore does not affect the carrier concentration and mobility measurements.

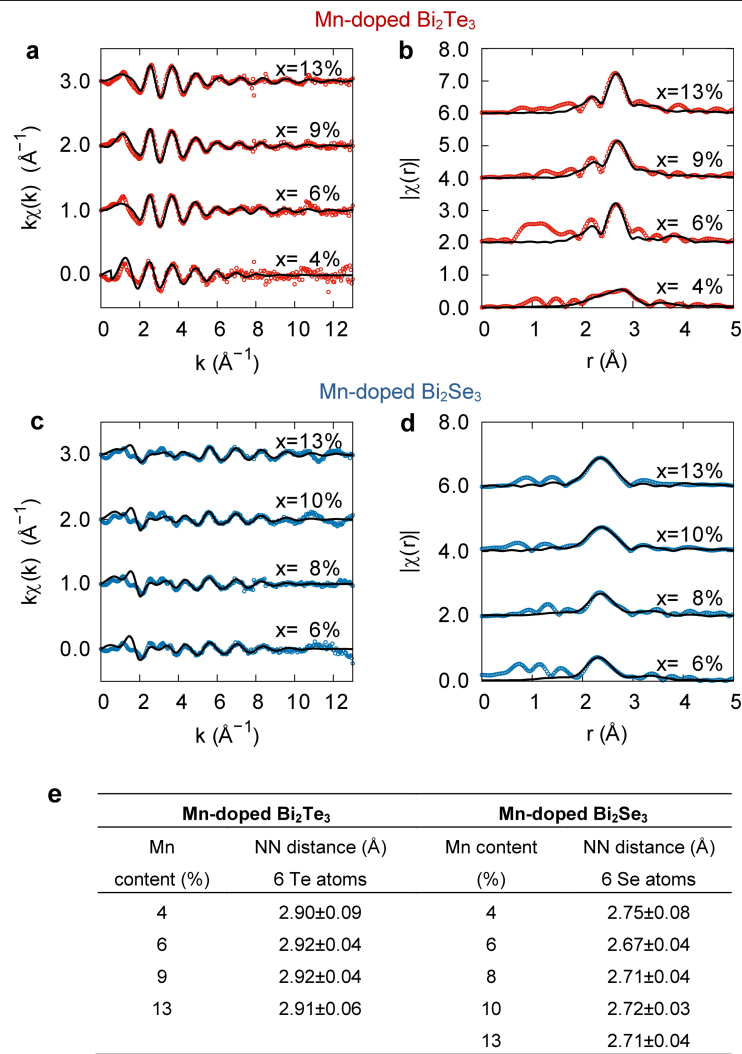


b

Mn Position	Mn-doped Bi ₂ Te ₃		Mn-doped Bi ₂ Se ₃	
	NN atoms	Distance (Å)	NN atoms	Distance (Å)
Center of septuple	6 Te	2.909	6 Se	2.719
Substitutional in QL (Bi)	3 Te(1)	3.073	3 Se(1)	2.972
	3 Te(2)	3.246	3 Se(2)	3.041
Octahedral in vdW gap	6 Te	2.858	6 Se	3.041
Tetrahedral in vdW gap	4 Te	2.533	4 Se	2.394
Te(1)/Se(1) sites next to vdW gap	3 Bi	3.073	3 Bi	2.972
	3 Te	3.673	3 Se	3.284
Te(2)/Se(2) sites at center of QL	6 Bi	3.246	6 Bi	3.041

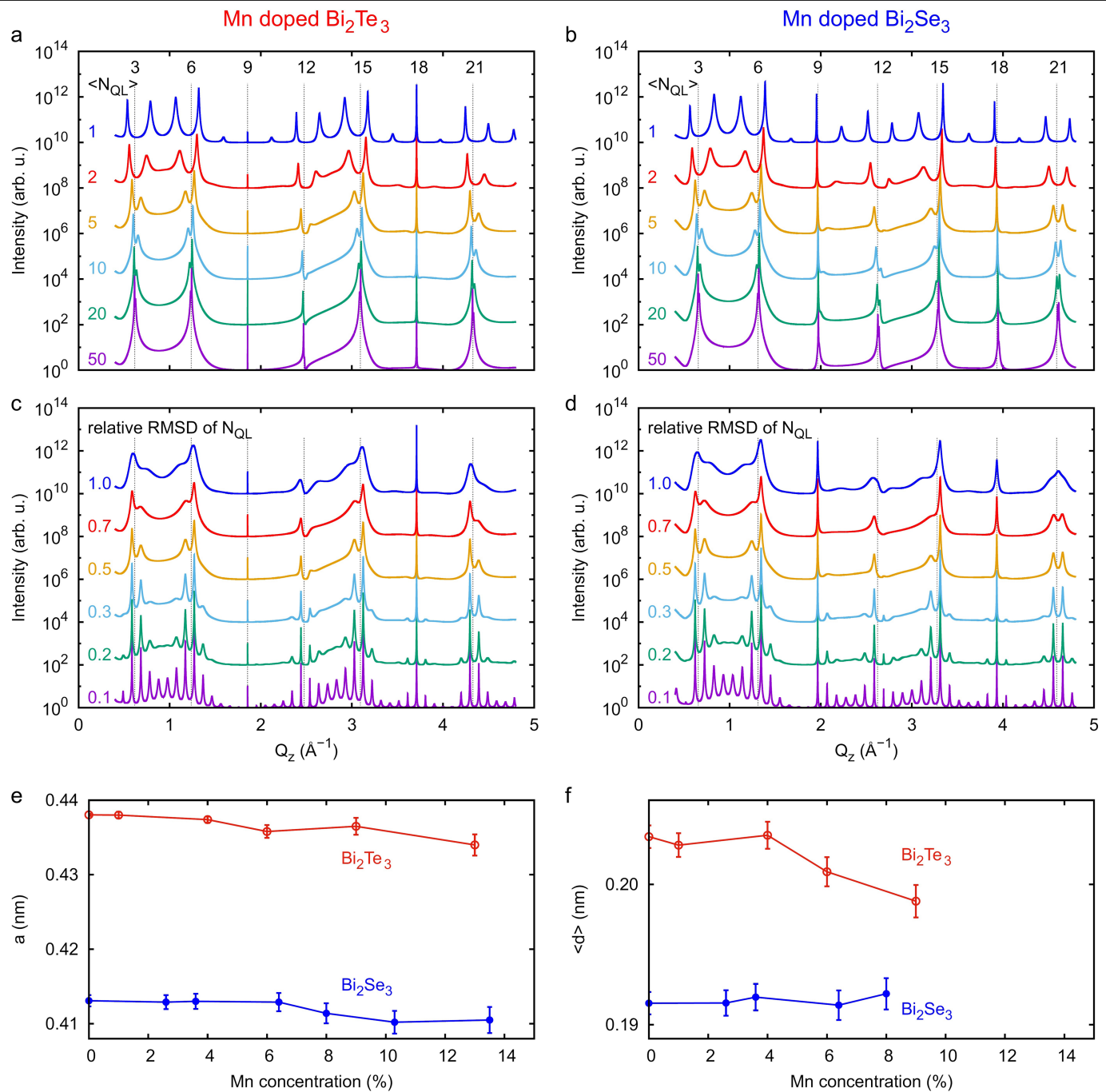
Extended Data Fig. 6 | Possible Mn-incorporation sites in Bi₂Te₃ and Bi₂Se₃.
a, Structure models for, left to right: Mn in the centre of septuple layers; Mn substituting for Bi in quintuple layers; and interstitial Mn in the van der Waals gaps on octahedral sites and tetrahedral sites. **b**, Nominal nearest-neighbour

(NN) distances of Mn atoms located in various positions as derived by EXAFS analysis, including also possible Mn on Te (Se) antisites in the quintuples. Index '1' refers to Te (Se) sites next to the van der Waals gaps, index '2' to those in the centre of the quintuple.



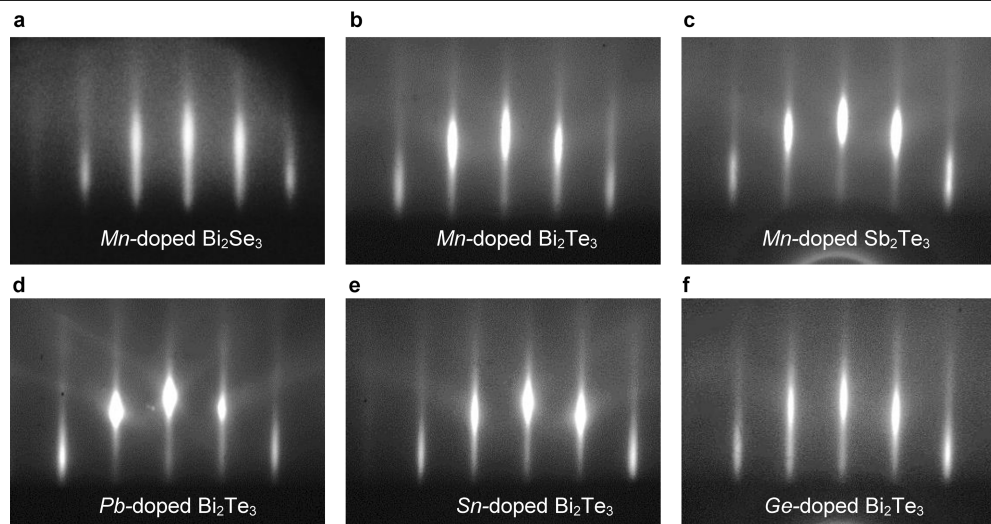
Extended Data Fig. 7 | Fits of EXAFS oscillations for different amounts of Mn doping. **a, b, Bi₂Te₃; **c, d**, Bi₂Se₃. Experimental data points are represented by the red or blue circles; black lines denote fitted curves (unlike the simulations in Fig. 3). **a, c** are plotted with respect to the wavevector **k** and the background-**

subtracted EXAFS absorption $\chi(k)$ is **k-weighted; **b, d** show the magnitude, that is, the absolute value of $\chi(R)$ after Fourier transformation. **e**, EXAFS-fitted nearest-neighbour distances of Mn atoms in Bi₂Te₃ and Bi₂Se₃ as a function of Mn concentration.**



Extended Data Fig. 8 | Simulated diffraction patterns. a–d, Varying paracrystal parameters of the septuple/quintuple heterostructures of: **a, c**, Mn-doped Bi_2Te_3 and **b, d**, Mn-doped Bi_2Se_3 . **a, b** depict the influence of different average numbers of quintuples, $\langle N_{\text{QL}} \rangle$, between the septuples with a fixed relative r.m.s. deviation (r.m.s.d.) of its distribution, set to 0.5. At high $\langle N_{\text{QL}} \rangle$, the system approaches the pure Bi_2Y_3 ($\text{Y} = \text{Te}, \text{Se}$) phase. **c, d** show simulations for different r.m.s.d. of the statistical distribution of the number of quintuples between the septuples but constant average separation, $\langle N_{\text{QL}} \rangle = 5$. The limit of r.m.s.d. = 0 corresponds to a perfectly periodic multilayer of five

quintuples alternating with one septuple; the additional maxima are the resulting superlattice satellite peaks. A larger r.m.s.d. means a more disordered multilayer. Dashed lines are plotted at positions of the (000 ℓ) peaks of the pure Bi_2Y_3 structure. **e** shows the average vertical (0001) lattice plane spacing (d) as a function of Mn concentration determined by the fit of the experimental diffraction spectra presented in Fig. 4a,b, and panel **f** the corresponding in-plane lattice constants determined from reciprocal space maps around the (101.20) reciprocal lattice point.

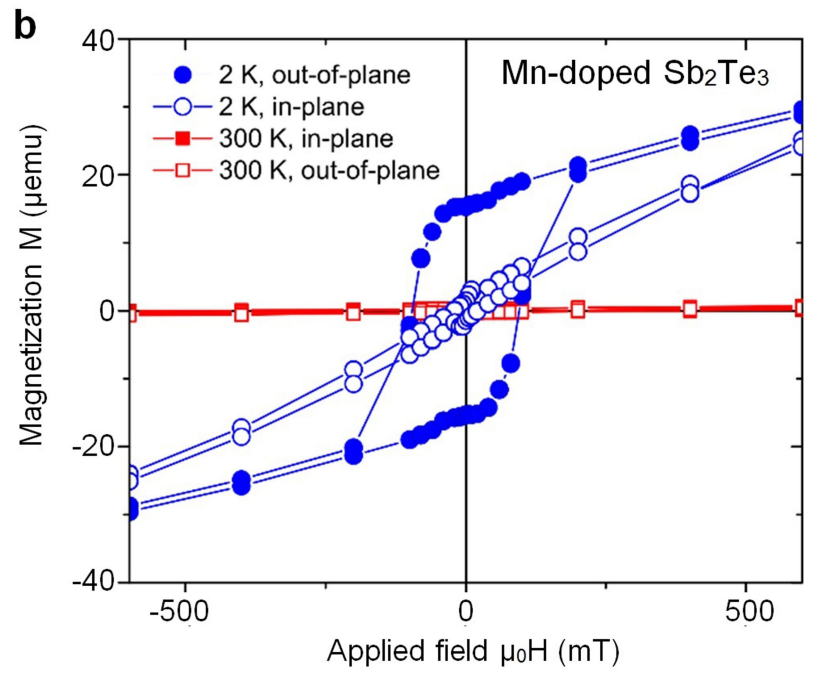
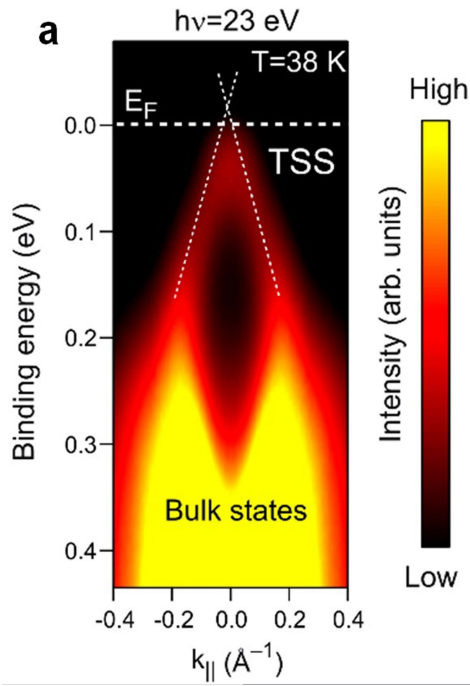


g

Material system	$\langle N_{QL} \rangle$	σ_{QL}	%SL
<i>Pb</i> - doped Bi ₂ Te ₃	2.1	1.3	14
<i>Sn</i> - doped Bi ₂ Te ₃	2.6	1.5	12
<i>Ge</i> - doped Bi ₂ Te ₃	3.0	1.5	10
<i>Mn</i> - doped Bi ₂ Te ₃	2.3	1.8	14
<i>Mn</i> - doped Sb ₂ Te ₃	3.5	1.7	11

Extended Data Fig. 9 | RHEED. **a–f**, Comparison of RHEED patterns for: **a** Mn-doped Bi₂Te₃; **b**, Mn-doped Bi₂Se₃; **c**, Mn-doped Sb₂Te₃; **d**, Pb-doped Bi₂Te₃; **e**, Sn-doped Bi₂Te₃; and **f**, Ge-doped Bi₂Te₃, recorded during epitaxial growth, showing perfect two-dimensional growth in all cases. The layer thickness was 200 nm and the dopant concentration was around 10% in all cases. The corresponding X-ray diffraction curves of the samples are shown in

Fig. 4, and the values derived from the fits using the septuple/quintuple paracrystal stacking model are listed in **g**. $\langle N_{QL} \rangle$ is the average number of quintuples between consecutive XBi_2Te_4 septuple layers; σ_{QL} is the relative r.m.s.d. of the distribution; and %SL is the occupancy of Mn sites in the septuple layers.



Extended Data Fig. 10 | Electronic and magnetic properties of Mn-doped Sb_2Te_3 . **a**, Angle-resolved photoemission spectrum recorded at a temperature of 38 K and a photon energy of 23 eV, showing p-type behaviour and that the Fermi level, E_F , is close to the Dirac point, the latter being only

slightly above the top of the valence band. **b**, In-plane and out-of-plane hysteresis curves, $M(H)$, recorded through SQUID at 2 K (blue) and 300 K (red), showing the same perpendicular magnetic anisotropy with easy axis normal to the surface as for Mn-doped Bi_2Te_3 (Fig. 2 and Extended Data Fig. 3).

Cooperative elastic fluctuations provide tuning of the metal–insulator transition

<https://doi.org/10.1038/s41586-019-1824-9>

G. G. Guzmán-Verri^{1,2,3*}, R. T. Brierley⁴ & P. B. Littlewood^{3,5*}

Received: 6 January 2017

Accepted: 11 September 2019

Published online: 18 December 2019

Metal-to-insulator transitions¹ driven by strong electronic correlations occur frequently in condensed matter systems, and are associated with remarkable collective phenomena in solids, including superconductivity and magnetism. Tuning and control of the transition holds the promise of low-power, ultrafast electronics², but the relative roles of doping, chemistry, elastic strain and other applied fields have made systematic understanding of such transitions difficult. Here we show that existing data^{3–5} on the tuning of metal-to-insulator transitions in perovskite transition-metal oxides through ionic size effects provides evidence of large systematic effects on the phase transition owing to dynamical fluctuations of the elastic strain, which have usually been neglected⁶. We illustrate this using a simple yet quantitative statistical mechanical calculation in a model that incorporates cooperative lattice distortions coupled to the electronic degrees of freedom. We reproduce the observed dependence of the transition temperature on the cation radius in the well studied manganite⁷ and nickelate⁸ materials. Because elastic couplings are generally strong, we anticipate that these conclusions will generalize to all metal-to-insulator transitions that couple to a change in lattice symmetry.

Metal-to-insulator transitions (MITs) driven by electronic correlations have energy scales of a few electronvolts, yet it is common to find that these phase transitions happen at temperatures corresponding to much lower energies¹. In the absence of a mechanism for fine-tuning the coupling constants, it is natural to look for entropic rather than enthalpic contributions with which to describe these transitions. Since all observed MITs couple to the lattice, one is then driven to look for phononic entropic contributions. As a hint at the origin of these interactions, a large number of transition-metal oxides with the ABO₃ perovskite crystal structure allow tuning of the MIT by not only by the choice and average valence of the electronically active B ion (usually a 3d transition metal) but also by the size of the electronically inactive A ion (usually a rare earth or alkaline earth ion)^{3,4,5,9}. This size effect can shift the transition temperature T_{MI} by hundreds of kelvin, and the widely accepted explanation¹⁴ is that the shift is due to a reduction in the electron bandwidth as the bond bending induced by ionic size changes the orbital overlap. However, the changes in bandwidth are not sufficiently large to explain such temperature variations^{10–13}. Moreover, it seems remarkable that a critical value of the ratio of interaction strength to bandwidth can be reached in every 3d transition-metal oxide, solely by varying the counterion¹⁴.

Instead, we propose here that even when the transition is clearly driven by local electronic correlations, anisotropic long-range forces induced by elastic compatibility conditions produce enormous entropic contributions to the free energy, which we show to be essential to a description of the variation of the MIT with cation size. We illustrate this with a model of highly fluctuating cooperative lattice

distortions that competes with a low-temperature phase of constant free energy, that is, a ferromagnetic metal for the manganites and a paramagnetic insulator for the nickelates. We do not aim to capture the complex charge, orbital and magnetic orderings of these materials, but rather the details of their high-temperature melted phase, where the entropy is dominated by the cooperative distortions. Our view is that the experiments in the manganite and nickelate series broadly implicate elastic interactions as being important for a wide class of MITs, not only in the perovskites.

In building our model, we account for the electronic degrees of freedom by assuming we can separate the energy into components that can be calculated locally while keeping the long-range physics explicit. At zero kelvin, state-of-the-art first-principles calculations can provide such local free energy, implicitly containing electron–phonon coupling on a unit cell as well as band structure energy and Coulomb correlation. We have not performed such calculations here. Instead, we have assumed that there is a simple functional outcome that can be parameterized, is the same across each material series, and is independent of the long-range energetic contribution.

Our approach has limitations: not every transition-metal oxide is electronically the same, for example, the bandwidth is not the only indicator or key parameter of structural changes in the electronic structure when varying the rare earth ion, nor are the local electronic correlations independent of the tolerance factor^{15,16}. These are idealizations which can describe real materials only approximately. Nonetheless, it allows us to illustrate that the non-trivial and subtle effects of long-range elastic interactions mediated between local degrees of freedom cannot

¹Centro de Investigación en Ciencia e Ingeniería de Materiales (CICIMA), Universidad de Costa Rica, San José, Costa Rica. ²Escuela de Física, Universidad de Costa Rica, San José, Costa Rica.

³Materials Science Division, Argonne National Laboratory, Argonne, IL, USA. ⁴Department of Physics, Yale University, New Haven, CT, USA. ⁵James Franck Institute, University of Chicago, Chicago, IL, USA. *e-mail: gian.guzman@ucr.ac.cr; pblittlewood@anl.gov

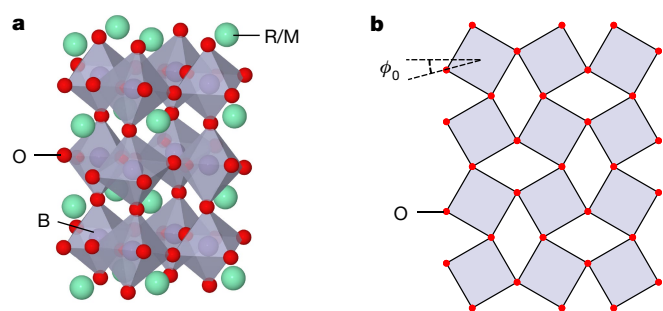


Fig. 1 | Perovskite lattices. **a**, Three-dimensional perovskite lattice showing the tilts of the BO_6 ($\text{B} = \text{Mn, Ni}$) octahedra. R is a rare earth element such as La, Pr, Nd and Sm; and M is an alkaline earth metal such as Ca, Sr and Ba. O is oxygen. **b**, Two-dimensional representation of the tilts used in our model, where ϕ_0 is an initial equilibrium antiferrodistortive rotation.

be ignored when it comes to determining the structural trends of MITs that couple to symmetry-breaking distortions.

The crystal structure of a perovskite transition-metal oxide consists of corner-sharing oxygen octahedra surrounding the B transition-metal ion, as shown in Fig. 1a. In general, the octahedra are tilted relative to their neighbours in an alternating pattern, and the tilt angle ϕ_0 increases with smaller A-site cation radius r_A . The large changes in functional behaviour of perovskites when varying ϕ_0 have led to proposals¹⁷ to engineer material properties by using a combination of strain, doping and pressure. In addition to variations of the atomic size, doping with A-site cations also introduces disorder in the cation size; careful distinction of the effects of doping and disorder for the manganites have demonstrated that disorder reduces the T_{MI} as effectively as varying r_A (ref.⁵).

Although purely electronic mechanisms to describe transition-metal oxides are appealing in their theoretical simplicity, it is known that the strong electron–phonon coupling means that the effects of lattice distortions cannot be neglected, and this is particularly well studied in manganites and nickelates^{18,19}. An electron that is localized by correlation effects in a unit cell will lower its energy further by the creation of a lattice distortion, which may have different symmetry in different materials. In the nickelates this is a simple breathing distortion, and in the manganites it is a so-called Jahn–Teller distortion, which lowers the cubic symmetry of the octahedron, as shown in Fig. 2a. The competition between this potential energy gain and the kinetic energy gained by delocalization to form a metal gives rise to the complex MIT phenomena in these materials.

The corner-sharing constraint on the octahedra introduces compatibility conditions between distortions at different lattice sites; when integrating out the phonon degrees of freedom these yield highly anisotropic, long-range interactions²⁰. Previous studies^{21,22} of phonon cooperativity in the manganites have explained the complex charge-ordered phases and mesoscopic structures that have been observed in the manganites and studied some effects of cooperative coupling on the transition²³. However, these studies did not consider the effect of octahedral tilting on the long-range interaction of the distortions. The purpose of this work is to study such effects, and in doing so, to construct a complete theory for cooperative elastic effects at a phase transition.

For illustration, we use a two-dimensional model of a perovskite, where we replace the octahedra with squares, as shown in Fig. 1b. Although the physics of bulk perovskites is three-dimensional, two-dimensional models^{21,22} of elastic interactions capture their anisotropy and long-range decay, which in turn have been shown to generate structural inhomogeneity over a wide range of length scales, which has been experimentally seen in transition-metal oxides and is relevant to our

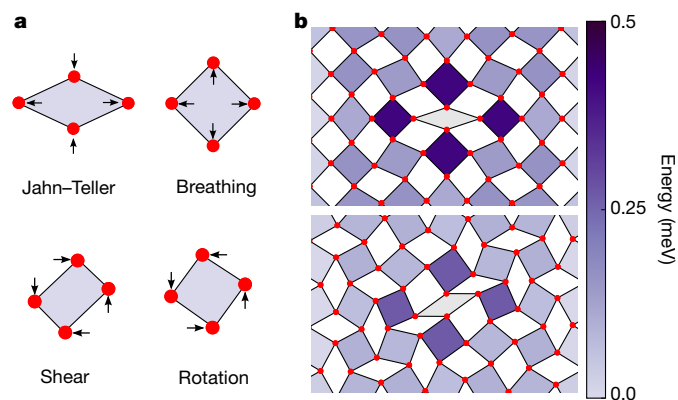


Fig. 2 | Lattice distortions and strain responses. **a**, Lattice distortions considered in our model. **b**, Strain responses of a lattice to a local Jahn–Teller distortion as a result of rotations. The colour of each square indicates the strain energy associated with the local distortions of that square. The grey parallelogram at the centre has a Jahn–Teller distortion of fixed amplitude. The strain fields weaken by allowing the BO_6 to tilt, as the energy is more effectively absorbed locally. Additional distortions on this site, such as shear, are allowed. Top, lattice with $\phi_0 = 0$. Bottom, lattice with $\phi_0 = 15^\circ$.

work. At a lattice site \mathbf{r} , the squares can undergo the distortions shown in Fig. 2a: deviatoric/Jahn–Teller modes T_r , dilatation/breathing modes D_r , shear modes S_r , and small rotations R_r of the squares from an initial equilibrium antiferrodistortive rotation ϕ_0 , that is, $\phi_r = (-1)^{|\mathbf{r}|} \phi_0 + R_r$. Assuming a harmonic energy penalty for creating distortions from an equilibrium configuration:

$$H = \sum_{\mathbf{r}} a_T T_r^2 + a_D D_r^2 + a_S S_r^2 \quad (1)$$

combined with the corner-sharing constraint, we can find an effective interaction $V_{rr'}(\phi_0)$ between different types of distortion, which gives rise to lattice cooperativity (see Supplementary Note 1). a_T , a_D and a_S are, respectively, the stiffness of the Jahn–Teller, breathing and shear distortions in a single, free octahedron and are independent of \mathbf{r} .

Figure 3 shows that the interaction strength is reduced by an increase tilt angle for Jahn–Teller distortions. This occurs because in the tilted configuration it is possible for the distortion to be accommodated by additional rotations to the neighbouring sites, rather than changes in the shape. Characteristic strain responses of the lattice to a local Jahn–Teller distortion with and without rotations are shown in Fig. 2b.

Both manganites⁷ and nickelates⁸ undergo first-order transitions from a characteristic low temperature phase to a high-temperature polaronic phase. This suggests that the motion of conduction electrons through the lattice is associated with the creation of local structural distortions that lead to a bad metal (highly resistive; $>10^{-5} \Omega \text{m}$)²⁴. When the distortion interaction $V_{rr'}(\phi_0)$ is reduced by changes in ϕ_0 , the high-temperature phase is favoured by a reduction in the polaron formation energy²³. To study this behaviour, we use $V_{rr'}(\phi_0)$ to form a statistical mechanical model for the distortions in this high-temperature phase, with a Hamiltonian:

$$H = \sum_{\mathbf{r}} \left[\frac{1}{2} \Pi_r^2 - \frac{\kappa}{2} Q_r^2 + \frac{\gamma}{4} Q_r^4 \right] + \sum_{\mathbf{r}, \mathbf{r}'} V_{rr'}(\phi_0) Q_r Q_{r'} - \sum_{\mathbf{r}} h_r Q_r \quad (2)$$

where Q_r is a Jahn–Teller (breathing) distortion for the manganites (nickelates) and Π_r its conjugate momentum. To model the compositional disorder that arises in the manganites from chemical substitution of the alkaline earth element at the A site of the perovskite structure, we consider a linear coupling of the lattice distortions Q_r to a local quenched random distortion h_r . We choose the h_r values to be normally distributed

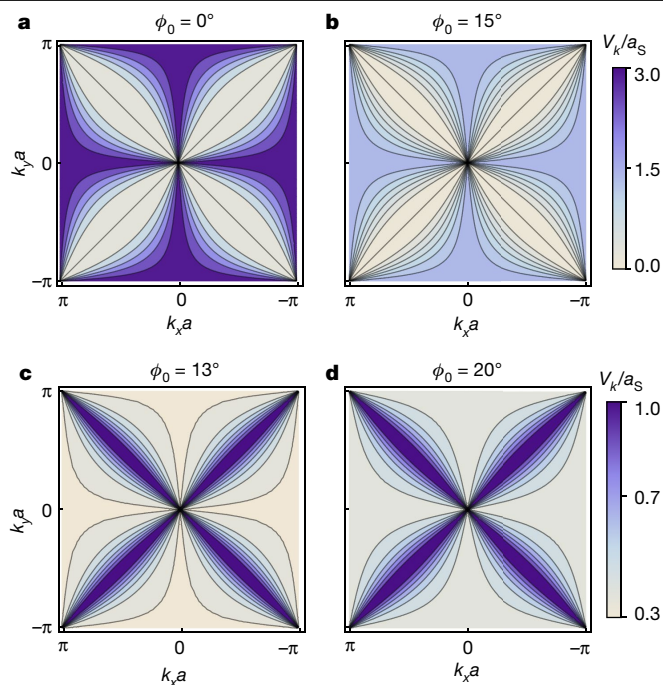


Fig. 3 | Effective elastic energy. **a, b**, Effective elastic energy for Jahn-Teller distortions in momentum space for $\phi_0 = 0^\circ$ (**a**) and $\phi_0 = 15^\circ$ (**b**). Rotations of the BO_6 octahedron allowed by the reduction of the A-cation size decrease the elastic energy. The characteristic ‘butterfly’ pattern is a consequence of the anisotropy and long-range nature of strain forces, which in turn can generate the salient nano- and meso-scale structural inhomogeneities, similar to those that have been observed in the manganites^{21,22} such as domain patterns in the form of stripes and tweeds formed by interwoven incommensurate structures. **c, d**, Effective elastic energy for breathing distortions in momentum space for $\phi_0 = 0^\circ$ (**c**) and $\phi_0 = 15^\circ$ (**d**). Tilts of the NiO_6 octahedron increase the effective elastic energy. The butterfly pattern is similar to that of the manganites, which produces structural inhomogeneity. (k_x, k_y) is a wavevector in the reciprocal space of the two-dimensional lattice of lattice constant a (shown in Fig. 1b).

with mean $\bar{h}_r = 0$ and variance $\bar{h}_r^2 = \Delta^2$. The negative sign of the Q_r^2 term describes the local tendency towards distortion due to the presence of electrons.

As described in the Methods and Supplementary Note 2, we use a variational approach to calculate the temperature, tilt angle and disorder

dependence of the free energy $F_{\text{lattice}}(T, \phi_0, \Delta)$ of Hamiltonian (2); and we identify the location of T_{MI} by comparing $F_{\text{lattice}}(T, \phi_0, \Delta)$ to a free energy $F_{\text{low}T}$ of the low-temperature ferromagnetic metal (paramagnetic insulator) phase of the manganites (nickelates). The results are shown in Fig. 4. Despite the over-simplicity of the model, the relationship between tilt angle, disorder and transition temperature is well reproduced. We do not attempt to describe the effects of the strain interactions on the MIT of the nickelates at low temperatures (see green region in Fig. 4a), because its magnetic ordering is different from that of the insulating phase above it. Similarly for the manganites, at low enough temperatures the polaronic, paramagnetic bad metal phase becomes either charge-ordered or glassy, beyond our approximations.

Here we have outlined a systematic theory for the incorporation of long-range elastic couplings into a simplified statistical mechanical theory of Mott-like phase transitions, where the electronic contributions to the free energy are incorporated at the level of Landau theory. That these elastic interactions are explicitly relevant for the manganites and the nickelates is confirmed by the ability of such a theory to systematically explain size effects or tolerance factor variations that have already been documented. However, the couplings, including their rough order of magnitude, are generic, and the ideas presented here will surely be relevant to other classes of materials such as the titanates⁹, high-temperature superconductors²⁵, ferroelectrics²⁶ and molecular fullerenes²⁷.

At low enough temperatures we also ought to consider other low-energy degrees of freedom such as spin fluctuations and electronic quantum fluctuations, which our model does not take into account. Doing so requires explicitly adding them to our model Hamiltonian and to our statistical mechanical solution through, for example, a variational scheme such as the Lang-Firsov transformation. Nonetheless, the model we employ does generate a quantum critical point attributable to elastic interactions alone. Moreover, the long range and anisotropy of these elastic couplings will modify the critical dynamics away from that arising from short-range models generated by purely electronic couplings.

We also note that our simple model provides an explanation for the observed tuning of the MIT under applied pressures. In both the manganites¹¹ and nickelates²⁸, hydrostatic compression decreases ϕ_0 . According to our model, this should result in an increase of T_{MI} promoted by the enhancement of the elastic interaction in the manganites, and vice versa for the nickelates. These are indeed the trends that have been observed in these materials^{28,29}. We believe a similar mechanism is at play when the transition is tuned with tensile and compressive stresses³⁰.

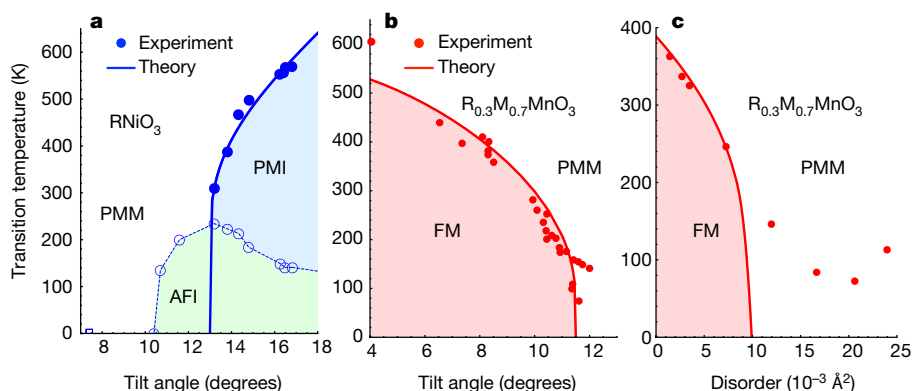


Fig. 4 | Comparison to experiments. **a**, Comparison of the nickelates for the transition temperature as a function of octahedral tilt angle. Filled and open circles are experimental transition temperatures⁸ for the paramagnetic insulator (PMI) and antiferromagnetic insulator (AFI) phases respectively. The extension of the green shading beyond the blue dashed line is an extrapolation. The open square (lower left) denotes LaNiO_3 , which is a polaronic, paramagnetic metal (PMM) at all temperatures. For the manganites, the

comparison is made with results⁵ (red circles) that separate the effect of tilt angle (**b**) and compositional disorder (**c**) on the transition from the paramagnetic, polaronic bad metal phase (PMM) to the ferromagnetic metal (FM) phase. Model parameters are fitted as described in the Methods. In Supplementary Note 3, we explore reasonable variations of the model parameters to demonstrate its generality.

The idea that cooperative phonon–phonon couplings tune the MIT is supported by a recent *ab initio* calculation¹⁹. By using density functional theory, it has been found that the tilts of the NiO_6 units in the nickelates destabilize their breathing distortions, which in turn are associated with the phase transition, thus providing a mechanism for tuning T_{MI} . However, density functional theory treats the elastic interactions only on average and it cannot produce finite temperature properties; T_{MI} has previously¹⁹ been obtained from fits to experiments with a Landau theory that has multiple sets of values for the model parameters depending on the tolerance factor. By contrast, we have calculated T_{MI} from a single set of model parameters, and the MIT is driven by entropic effects that result from elastic couplings, thus providing a physical interpretation of the *ab initio* results.

We conclude by noting that the good agreement we found in these two systems suggests that our fundamental assumption that energy can be separated into a relatively simple local free energy plus a complex long-range energetic contribution could provide a basis for a fully computational methodology that could be applied relatively simply to very complex oxides in general.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1824-9>.

- Imada, M., Fujimori, A. & Tokura, Y. Metal-insulator transitions. *Rev. Mod. Phys.* **70**, 1039 (1998).
- Yang, Z., Ko, C. & Ramanathan, S. Oxide electronics utilizing ultrafast metal-insulator transitions. *Annu. Rev. Mater. Res.* **41**, 337–367 (2011).
- Torrance, J. B., Lacorre, P., Nazzari, A. I., Ansaldo, E. J. & Niedermayer, Ch. Systematic study of insulator-metal transitions in perovskites RNiO_3 ($R = \text{Pr, Nd, Sm, Eu}$) due to closing of charge-transfer gap. *Phys. Rev. B* **45**, 8209–8212 (1992).
- Hwang, H. Y., Cheong, S.-W., Radaelli, P. G., Marezio, M. & Batlogg, B. Lattice effects on the magnetoresistance in doped LaMnO_3 . *Phys. Rev. Lett.* **75**, 914–917 (1995).
- Rodríguez-Martínez, L. M. & Attfield, J. P. Cation disorder and size effects in magnetoresistive manganese oxide perovskites. *Phys. Rev. B* **54**, R15622(R) (1996).
- Khomskii, D. I. *Transition Metal Compounds* (Cambridge University Press, 2014).
- Tokura, Y. Critical features of colossal magnetoresistive manganites. *Rep. Prog. Phys.* **69**, 797–851 (2006).
- Catalano, S. et al. Rare-earth nickelates RNiO_3 : thin films and heterostructures. *Rep. Prog. Phys.* **81**, 046501 (2018).
- Katsufuji, T., Taguchi, Y. & Tokura, Y. Transport and magnetic properties of a Mott–Hubbard system whose bandwidth and band filling are both controllable: $\text{R}_{1-x}\text{Ca}_x\text{TiO}_{3+y/2}$. *Phys. Rev. B* **56**, 10145–10153 (1997).

- Sarma, D. D., Shanthi, N. & Mahadevan, P. Electronic structure and the metal-insulator transition in LnNiO_3 ($\text{Ln} = \text{La, Pr, Nd, Sm}$ and Ho): bandstructure results. *J. Cond. Matt. Phys.* **6**, 10467–10474 (1994).
- Radaelli, P. G. et al. Structural effects on the magnetic and transport properties of perovskite $\text{A}_{1-x}\text{A}_x\text{MnO}_3$ ($x = 0.25, 0.30$). *Phys. Rev. B* **56**, 8265–8276 (1997).
- Medarde, M., Lacorre, P., Conder, K., Fauth, F. & Furrer, A. Giant ^{16}O – ^{18}O isotope effect on the metal-insulator transition of RNiO_3 perovskites ($R = \text{rare earth}$). *Phys. Rev. Lett.* **80**, 2397–2400 (1998).
- Varignon, J., Grisolia, M. N., Íñiguez, J., Barthélémy, A. & Bibes, M. Complete phase diagram of rare-earth nickelates from first-principles. *npj Quant. Mater.* **2**, 21 (2017).
- Fujimori, A. Electronic structure of metallic oxides: band-gap closure and valence control. *J. Phys. Chem. Solids* **53**, 1595–1602 (1992).
- Pavarini, E., Yamasaki, A., Nuss, J. & Andersen, O. K. How chemistry controls electron localization in $3d^1$ perovskites: a Wannier-function study. *New J. Phys.* **7**, 188 (2005).
- Han, Q. & Millis, A. Lattice energetics and correlation-driven metal-insulator transitions: the case of Ca_2RuO_4 . *Phys. Rev. Lett.* **121**, 067601 (2018).
- Rondinelli, J. M., May, S. J. & Freeland, J. W. Control of octahedral connectivity in perovskite oxide heterostructures: an emerging route to multifunctional materials discovery. *MRS Bull.* **37**, 261–270 (2012).
- Millis, A. J., Littlewood, P. B. & Shraiman, B. I. Double exchange alone does not explain the resistivity of $\text{La}_{1-x}\text{Sr}_x\text{MnO}_3$. *Phys. Rev. Lett.* **74**, 5144–5147 (1995).
- Mercy, A. & Bieder, J., Íñiguez, J. & Ghosez, P. Structurally triggered metal-insulator transition in rare-earth nickelates. *Nat. Commun.* **8**, 1677 (2017).
- Kartha, S., Krumhansl, J. A., Sethna, J. P. & Wickham, L. K. Disorder-driven pretransitional tweed pattern in martensitic transformations. *Phys. Rev. B* **52**, 803 (1995).
- Ahn, K. H., Lookman, T. & Bishop, A. R. Strain-induced metal-insulator phase coexistence in perovskite manganites. *Nature* **428**, 401–404 (2004).
- Ahn, K. H., Seman, T. F., Lookman, T. & Bishop, A. R. Role of complex energy landscapes and strains in multiscale inhomogeneities in perovskite manganites. *Phys. Rev. B* **88**, 144415 (2013).
- Millis, A. J. Cooperative Jahn–Teller effect and electron-phonon coupling in $\text{La}_{1-x}\text{A}_x\text{MnO}_3$. *Phys. Rev. B* **53**, 8434–8441 (1996).
- Jaramillo, R. et al. Origins of bad-metal conductivity and the insulator–metal transition in the rare-earth nickelates. *Nat. Phys.* **10**, 304–307 (2014).
- Attfield, J. P., Kharlanov, A. L. & McAllister, J. A. Cation effects in doped La_2CuO_4 superconductors. *Nature* **394**, 157–159 (1998).
- Balachandran, P. V., Broderick, S. R. & Rajan, K. Identifying the inorganic gene for high-temperature piezoelectric perovskites through statistical learning. *Proc. R. Soc. A* **467**, 2271–2290 (2011).
- Zadik, R. H. et al. Optimized unconventional superconductivity in a molecular Jahn–Teller metal. *Sci. Adv.* **1**, e1500059 (2015).
- Obradors, X. et al. Pressure dependence of the metal-insulator transition in the charge-transfer oxides RNiO_3 ($R = \text{Pr, Nd, Nd}_{0.7}\text{La}_{0.3}$). *Phys. Rev. B* **47**, 12353–12356 (1993).
- Fontcuberta, J., Laukhin, V. & Obradors, X. Local disorder effects on the pressure dependence of the metal-insulator transition in manganese perovskites. *Appl. Phys. Lett.* **72**, 2607–2609 (1998).
- Liu, J. et al. Heterointerface engineered electronic and magnetic phases of NdNiO_3 thin films. *Nat. Commun.* **4**, 2714 (2013).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2019

Methods

Statistical mechanical solution

We use a variational pair-distribution function that incorporates mean-field behaviour, Gaussian corrections to the thermal and quantum fluctuations, and averaging over compositional disorder at the level of the replica method³¹. Details are provided in Supplementary Note 2.

Model parameters

Our model has six parameters ($\kappa, \gamma, a_D, a_S, a_T$ and $F_{\text{low}T}$), which are reduced to five because a_T (a_D) is combined with κ for the manganites (nickelates). We begin by choosing a set of physically reasonable parameters that give phonon frequencies that are in order-of-magnitude agreement with the observed relevant modes^{32,33}. We then take the resulting set of parameters and fine-tune them to fit the observed dependence of T_{MI} with the tolerance factor and compositional disorder: $F_{\text{low}T}$ is a parameter of the model assumed to be independent of T, ϕ_0 and Δ , fixed by the observed onset of the MIT, that is, $F_{\text{lattice}}(T=0\text{ K}, \phi_0=\phi_{\text{onset}}, \Delta=0)=F_{\text{low}T}$, where $\phi_{\text{onset}} \approx 11.5^\circ$ (12.5°) for the manganites (nickelates). The dependence of T_{MI} on ϕ_0 shown in Fig. 4a, b is given by $F_{\text{lattice}}(T_{\text{MI}}, \phi_0, \Delta=0)=F_{\text{low}T}$, while the dependence of T_{MI} on Δ shown in Fig. 4c is given by $F_{\text{lattice}}(T_{\text{MI}}, \phi_0=80^\circ, \Delta)=F_{\text{low}T}$ and by rescaling Δ by a constant factor (α) to match the units of cation variance. The resulting values are given in Extended Data Table 1.

Data availability

Requests for materials should be addressed to G.G.G.-V., and P.B.L.

31. Guzmán-Verri, G. G., Littlewood, P. B. & Varma, C. M. Paraelectric and ferroelectric states in a model for relaxor ferroelectrics. *Phys. Rev. B* **88**, 134106 (2013).
32. Zaghrioui, M., Bulou, A., Lacorre, P. & Laffez, P. Electron diffraction and Raman scattering evidence of a symmetry breaking at the metal-insulator transition of NdNiO_3 . *Phys. Rev. B* **64**, 081102 (2001).
33. Martín-Carrón, L., de Andrés, A., Martínez-Lope, M. J., Casais, M. T. & Alonso, J. A. Raman phonons as a probe of disorder, fluctuations, and local structure in doped and undoped orthorhombic and rhombohedral manganites. *Phys. Rev. B* **66**, 174303 (2002).

Acknowledgements We acknowledge discussions with G. Lonzarich, H. Park and F. Ballar-Trigueros. Work at Argonne National Laboratory is supported by the US Department of Energy, Materials Science Division, Office of Basic Energy Sciences under contract number DE-AC02-06CH11357. G.G.G.-V. acknowledges support from the Vice-rectory for Research (project number 816-B7-601), and the Office of International Affairs at the University of Costa Rica, the Royal Society International Exchanges programme (grant number IES\R3\170025), Churchill College (University of Cambridge). G.G.G.-V. thanks the Department of Materials Science and Metallurgy and the Cavendish Laboratory at the University of Cambridge (where part of this work was done) for hospitality. R.T.B. acknowledges support from the Yale Prize Postdoctoral Fellowship and Homerton College (University of Cambridge).

Author contributions P.B.L. conceived the study. G.G.G.-V. and R.T.B. performed the calculations. All authors constructed the model, wrote the manuscript, discussed the results and implications at all stages.

Competing interests R.T.B. is currently an editor at *Nature Communications*.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1824-9>.

Correspondence and requests for materials should be addressed to G.G.G.-V. or P.B.L.

Peer review information *Nature* thanks Mona Berciu, Paolo Radaelli and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Extended Data Table 1 | Model parameters

	κ [meV ²]	γ [meV ³]	a_D [meV ²]	a_S [meV ²]	a_T [meV ²]	$F_{\text{low } T}$ [meV]	α [Å meV ⁻¹]
(R _{0.3} M _{0.7})MnO ₃	3.2×10^3	1.8×10^5	2.8×10^4	9.5×10^3	combined with κ	38	4.71
RNiO ₃	12.1×10^3	13.3×10^5	combined with κ	48.3×10^3	1.21×10^4	64	-

R is a rare earth element such as La, Pr, Nd, and Sm; and M is an alkaline earth metal such as Ca, Sr and Ba.

Molecular heterogeneity drives reconfigurable nematic liquid crystal drops

<https://doi.org/10.1038/s41586-019-1809-8>

Wei-Shao Wei^{1,2*}, Yu Xia^{2,3}, Sophie Ettinger^{1,2}, Shu Yang^{2,3} & A. G. Yodh^{1,2}

Received: 19 April 2019

Accepted: 25 September 2019

Published online: 18 December 2019

With few exceptions^{1–3}, polydispersity or molecular heterogeneity in matter tends to impede self-assembly and state transformation. For example, shape transformations of liquid droplets with monodisperse ingredients have been reported in equilibrium^{4–7} and non-equilibrium studies^{8,9}, and these transition phenomena were understood on the basis of homogeneous material responses. Here, by contrast, we study equilibrium suspensions of drops composed of polydisperse nematic liquid crystal oligomers (NLCs). Surprisingly, molecular heterogeneity in the polydisperse drops promotes reversible shape transitions to a rich variety of non-spherical morphologies with unique internal structure. We find that variation of oligomer chain length distribution, temperature, and surfactant concentration alters the balance between NLC elastic energy and interfacial energy, and drives formation of nematic structures that range from roughened spheres to ‘flower’ shapes to branched filamentous networks with controllable diameters. The branched structures with confined liquid crystal director fields can be produced reversibly over areas of at least one square centimetre and can be converted into liquid crystal elastomers by ultraviolet curing. Observations and modelling reveal that chain length polydispersity plays a crucial role in driving these morphogenic phenomena, via spatial segregation. This insight suggests new routes for encoding network structure and function in soft materials.

We study NLC drops, tens of micrometres in diameter, dispersed in water containing sodium dodecyl sulfate (SDS) surfactant. SDS creates a strong preference for homeotropic anchoring, wherein the nematic director (molecular orientation) is perpendicular to the drop surface (Fig. 1a). Each surfactant-stabilized NLC emulsion drop contains a mixture of 1,4-bis-[4-(6-acryloyloxyhexyloxy)benzoyloxy]-2-methylbenzene (RM82) monomers¹⁰ and RM82 oligomers with variable chain lengths (Fig. 1a, details in Extended Data Fig. 1a). Importantly, compared to small-molecule liquid crystals, chain length polydispersity of the NLCs offers new degrees of freedom which can profoundly affect drop morphology, and the NLCs are readily crosslinked to lock-in nematic order and morphology.

Essential features of the shape transitions are shown in Fig. 1. At a high temperature (around 90 °C), while in the nematic phase, the drop is spherical (Fig. 1a, c). Here, interfacial energy is large compared to bulk director elastic energy. On cooling, the surface tension and bulk elasticity vary, and excess interface is created. This interplay destabilizes the drop, facilitating spontaneous polymorphic transitions to non-spherical equilibrium structures (see exemplar images in Fig. 1c–h; director configurations in Fig. 1a, b; Supplementary Video 1). These polymorphic shape transitions are reversible and repeatable via temperature cycling (see Extended Data Fig. 2, Supplementary Video 2).

Arguably, the major factor driving these shape transitions is the interfacial tension, γ , which decreases with decreasing temperature. This trend was confirmed in pendant drop experiments using large,

millimetre-size NLC drops (Extended Data Fig. 3), and has been reported in other systems^{7,11}. The decrease in γ is believed to arise from greater molecular ordering at the interface^{7,11,12}. In our case, surfactant-induced anchoring and the stronger alignment of NLCs at lower temperatures enhance interfacial ordering. Additionally, NLC elastic constants increase with decreasing temperature¹³. Thus, with reduction of γ , anisotropic bulk elasticity becomes more important for determining drop shape, increasing the potential for spontaneous shape transitions⁴.

To elucidate the consequences of these effects, we constructed a morphology ‘state’ diagram from samples that experienced the same heating–cooling cycles. Figure 2a–d exhibits the repeatable drop morphologies: smooth/roughened spheres, flowers, large- and small-diameter (d) filament networks. The state diagram (Fig. 2e) is shown as a function of surfactant concentration and mean oligomer chain length, $\langle \ell \rangle$, which increases with total oligomerization time.

Consider the effects of NLC average chain length. Though all samples are polydisperse, samples prepared with longer oligomerization processing times contain a larger ratio of long-chain to short-chain oligomers and have longer $\langle \ell \rangle$ (Fig. 2f, Methods, Extended Data Fig. 4). Larger NLC elastic constants are expected¹⁴ for longer $\langle \ell \rangle$. Additionally, interfacial oligomer assembly with longer $\langle \ell \rangle$ tends to reduce γ (see Extended Data Fig. 3); longer NLCs increase oligomer chain hydrophobicity, which induces more interfacial ordering^{15,16}. As temperature is lowered, these trends favour drop morphologies with large surface

¹Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA, USA. ²Laboratory for Research on the Structure of Matter (LRS), University of Pennsylvania, Philadelphia, PA, USA. ³Department of Materials Science and Engineering, University of Pennsylvania, Philadelphia, PA, USA. *e-mail: weiwe@sas.upenn.edu

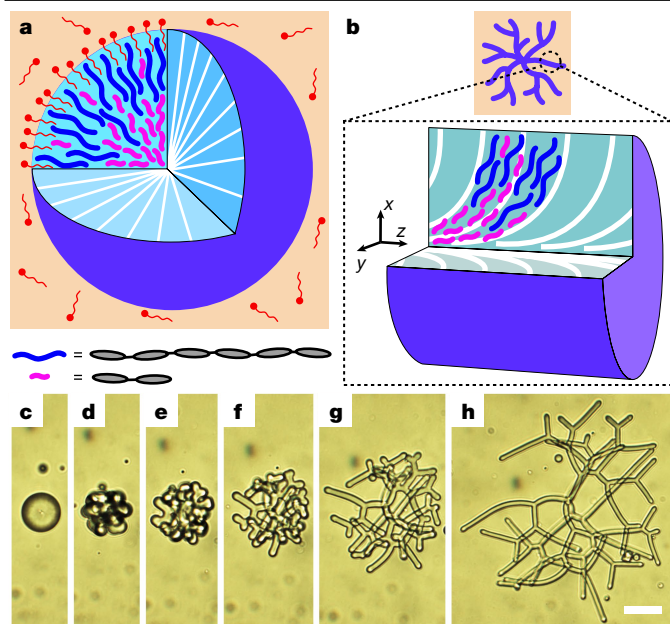


Fig. 1 | Spontaneous shape transition of NLCO drops. **a, b**, Schematics depicting a surfactant-stabilized NLCO sphere (**a**) and filament (**b**) shown in purple in a background aqueous phase (orange). The red symbols in **a** represent surfactant molecules. White lines represent the nematic director. The concept of spatial segregation by oligomer chain length is illustrated. The lower inset in **a** shows how, for simplicity, long-chain/short-chain species are represented by blue/pink solid curves representing hexamers and dimers, respectively (grey ellipses represent monomers). The inset in **b** shows a magnified view of a filament cross-section. The z axis in **b** is parallel to (and coincident with) the central filament axis. **c–h**, An NLCO drop spontaneously (and reversibly) evolves from a sphere to an extended and branched filamentous drop as the temperature is reduced from 90 °C (**c**) to 20 °C (**h**). Scale bar in **h** (for **c–h**), 20 μm .

area. More quantitatively, our measurements indicate a crossover transition to non-sphericity when $\langle \ell \rangle \approx 1.5$ times the monomer length in 0.1 wt% SDS solutions.

Other factors modify surface tension and affect the state diagram. In contrast to work on non-equilibrium liquid crystal filament formation^{8,9}, our experiments were largely carried out below the critical micelle concentration (CMC) of SDS wherein increasing SDS concentration lowers γ and helps facilitate interfacial roughening and filament formation (Fig. 2e). Additional experiments confirmed that primary phenomena are not unique to SDS surfactant, nor limited to concentrations below the CMC (Extended Data Figs. 5, 6).

Given this emerging qualitative picture, we sought a more quantitative understanding. To this end, we determined NLCO director configurations using polarization optical microscopy (POM) with a full-wave retardation plate. We then applied simple models to determine their free energies. Here we focus on the two limiting morphologies: spherical versus filamentous drops. Spheres exhibit the classic POM texture corresponding to a radial ‘hedgehog’ director configuration with a central topological charge of +1 (that is, a point defect; Fig. 3a). Filaments exhibit an escaped-radial director configuration (Fig. 3b, c). For all geometries, the single point defect always remains at the drop centre (Fig. 3a, d, e).

To predict shape transition phenomena, modelling must account for energetics associated with both interfacial tension and elastic bulk free energy of the confined NLCOs. The calculations require knowledge about the splay (K_{11}), twist (K_{22}), bend (K_{33}) and saddle-splay (K_{24}) NLCO elastic constants, as well as interface anchoring strength. We utilize well

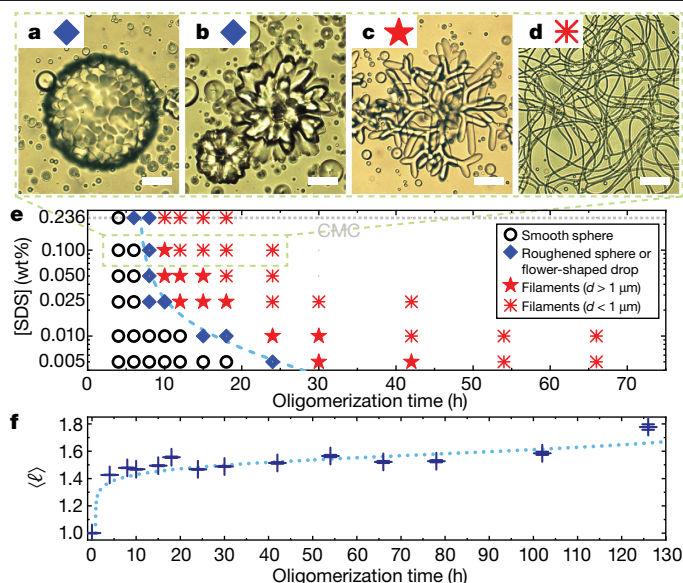


Fig. 2 | Equilibrium NLCO drop morphologies as a function of mean oligomer chain length and surfactant concentration. **a–d**, Optical images of common NLCO drop morphologies at room temperature in a 0.1 wt% SDS solution. With increasing mean oligomer chain length $\langle \ell \rangle$, the morphologies shown are: **a**, roughened spheres (see also Fig. 4a); **b**, flower-shaped drops; and **c, d**, filamentous structures with decreasing filament diameter. **e**, A state diagram, constructed from data, exhibits the equilibrium morphology (examples in **a–d**; key shows data symbols for each morphology) versus $\langle \ell \rangle$ and surfactant concentration. Increasing either parameter facilitates filament formation. The dashed curve separates spherical and non-spherical drop regimes. Experiments were carried out below the critical micelle concentration (CMC), approximately 0.236 wt% SDS. **f**, Size exclusion chromatography shows that $\langle \ell \rangle$ (in units of monomer length) increases with oligomerization processing time (horizontal axis of plots in **e** and **f**). Bars indicate the spread in $\langle \ell \rangle$, which mainly arises from our inability to detect the longer-chain components. Dotted curve is a guide for the eye. Scale bar in **a–d**, 20 μm .

known models for the sphere with radial hedgehog director configuration¹⁷ and for cylinders with escaped-radial director configuration¹⁸. Importantly, all models assume a homogeneous, monodisperse chain length distribution for the NLCOs. The models employ γ from pendant drop experiments, measured drop dimensions, and estimates of elastic constants and interfacial anchoring strength based on small-molecule liquid crystals¹⁹. Calculations (see Methods, Extended Data Table 1) yield conditions for the sphere–filament free energy instability.

Surprisingly, we found that free energy instabilities only occur if our system has either an unphysically large saddle-splay elastic constant ($K_{24} > 30K_{11}$), or a value of γ that is much smaller (about one-tenth) than those obtained in the pendant drop experiments. Given this discrepancy, we were compelled to consider other effects. In the process, we uncovered the importance of NLCO chain length polydispersity (molecular heterogeneity), which can profoundly affect interface and core energetics.

First, consider elastic free energy gradients in small (about 10–30 μm) drops versus the large (millimetre-size) drops used to obtain γ . In small drops, the nematic adopts a single-domain configuration in which the director field experiences large splay distortion near drop centre for spheres, or large splay and bend distortion near the central axis for filaments (see Methods). The millimetre-size drops, by contrast, are composed of many nematic domains, each with different director orientation, and they are filled with many disclination lines. Although director fields in the large drops can be locally non-uniform, most microdomains remain uniformly aligned, and splay/bend elasticity gradients are not large. Thus, the overall elastic gradient effects in millimetre-size drops are small.

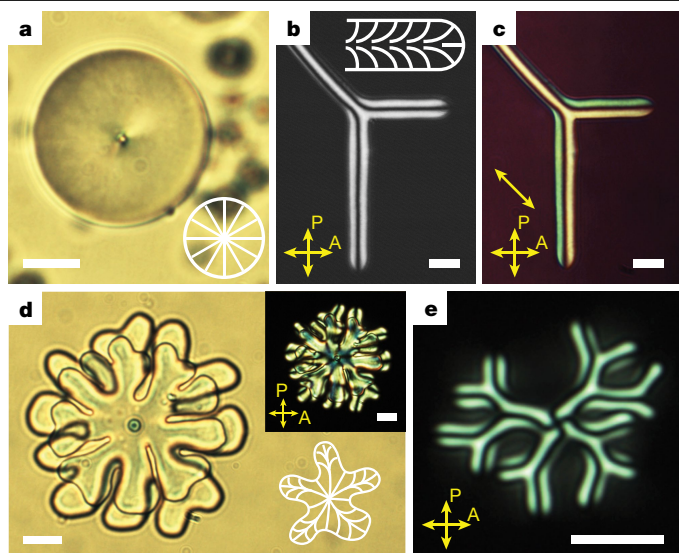


Fig. 3 | Director configurations within NLCO structures. **a**, Bright-field image of a spherical NLCO drop with a single, central radial 'hedgehog' defect, shown schematically at bottom right. **b, c**, POM (**b**) coupled with a full-wave retardation plate (**c**) enables identification of nematic director configurations within a section of a y-branched NLCO filamentous structure. The crossed double-headed arrows represent the pass axes of the polarizer (P) and analyser (A), respectively; the diagonal double-headed arrow indicates an inserted full-wave retardation plate. In **c**, yellow indicates northwest–southeast alignment; cyan indicates northeast–southwest alignment. Inset in **b**, 2D projected director field schematic. **d**, Bright-field image of a flower-shaped structure with a hedgehog defect, shown schematically at bottom right. Inset shows the corresponding POM image. **e**, The POM image of a filamentous structure with a hedgehog appearing as the cross-pattern. We note that the hedgehog defect always remains at the centre of the drop, no matter how complex the structure is. The structures shown in **d** and **e** were weakly confined in a quasi-2D chamber such that filaments tended to grow perpendicularly to the optical axis, providing easy observation of the point defect. Scale bar in **a–e**, 10 μm .

In the micrometre-size drops and cylinders, however, oligomers experience strong chain-length-dependent driving forces. These forces are induced by elastic energy gradients of the director within the drop, that is, higher elastic energy density near the core. To reduce system

elastic energy, short chains, which have smaller associated elastic constants, will migrate towards the core, and long chains, which have larger associated elastic constants, will migrate towards the drop surface (Fig. 1a, b). A simple two-component macromer–monomer demixing model with a mean oligomer chain length of $\langle \ell \rangle \approx 1.7$ for the outer shell and a mean oligomer chain length in the core of $\langle \ell \rangle \approx 1.39$, for example, predicts around a 10% decrease in elastic free energy due to reduced average rod length in the core.

Second, as a result of these elastic free energy gradients, interfacial tension in the small drops is lowered greatly and will differ from that in large drops. Stronger nematic ordering at the interface driven by the greater numbers of long-chain oligomers will decrease γ by a factor of 10 or more compared to drops with homogeneous oligomer distributions (see Extended Data Fig. 3). Thus, spatial redistribution of polydisperse oligomers in the small drops resolves the issues raised by the original homogeneous model. A conservative reduction of γ by ten times to approximately 0.1 mN m^{-1} yields sphere–filament instabilities with reasonable saddle-splay moduli. Importantly, our experiments with bidisperse distributions of monomers and macromers, made by a different chemical procedure¹⁰, support this central new insight about the effects of NLCO polydispersity. (See Methods for these latter experiments.) Previously, shape transitions were observed in molecularly heterogeneous systems²⁰, and the phenomenon of segregation by size, broadly defined, has been reported in lipid membrane systems²¹, near topological defects in liquid crystal simulations²², and in liquid crystal polymer systems mediated by size-dependent nematic/isotropic transition temperatures²³. It is perhaps useful to reconsider these phenomena in light of our experiments.

By contrast, as noted above, spatial redistribution of oligomers is not anticipated in millimetre-size drops. Segregation is driven largely by director distortion near uniformly distributed disclinations and domain walls and will not induce overall migration to the surface. Thus, the surface of the millimetre-size drop is likely to remain a uniform mixture of oligomers with mean chain length set by the initial NLCO distribution. The pendant millimetre-drop γ therefore reflects 'average behaviour' of the mixture; it should be much larger than the γ of micrometre-size drops. Possible curvature-induced forces driving segregation will also be less important in large versus small drops.

To harvest these self-assembled morphologies for new functional materials, the NLCOs were ultraviolet-crosslinked into nematic liquid crystal elastomers (NLCEs). Representative scanning electron

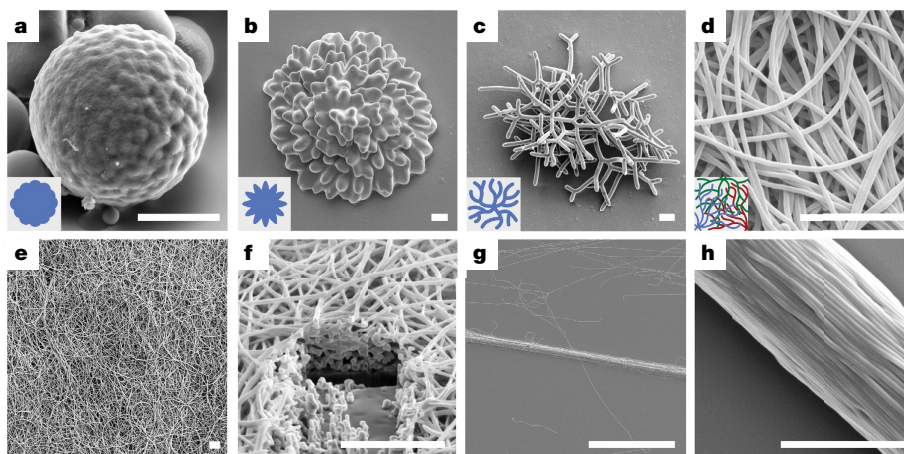


Fig. 4 | SEM images of self-assembled NLCEs. **a–d**, Crosslinked NLCEs: **a**, roughened sphere; **b**, flower; and **c, d**, filamentous structures with decreasing diameter. (See Fig. 2a–d for corresponding NLCO morphologies observed in bright-field microscopy.) The objects shown in **a–c** evolved from a single drop and are depicted schematically in the insets; **d** shows thin intertwined filamentous structures grown from several drops (inset shows

filaments with different colours grown from different drops). **e, f**, Partial view of a centimetre-wide free-standing NLCE fibrous mat (**e**) and its cross-section (**f**; cut by a focused Ga ion beam). See Extended Data Fig. 7 for a macroscopic image of the mat. **g, h**, Oriented NLCE fibre yarn at different magnifications. Scale bars: **a–f, h**, 5 μm ; **g**, 100 μm .

microscopy (SEM) images of the NLCEs are in Fig. 4a–d. By comparing POM images of corresponding NLCE and NLCO structures, we confirmed that director configurations are well-maintained.

Last, we show that these NLCE fibres can be densely packed into centimetre-wide and few-micrometres-thick, non-woven, free-standing NLCE mats (Fig. 4e, f) by sedimentation. These structures could find applications in responsive filtration and smart fabrics. A NLCE yarn consisting of well-aligned fibres (Fig. 4g, h) was made by directly pulling fibres out of aqueous solution; these yarn-like objects could find use in artificial muscles^{24,25} or tunable waveguides. Actuation with the current fibres is not yet apparent in our preliminary studies, probably because of high crosslinking density due to insufficient oligomerization. Variation of liquid crystal oligomer chemistry²⁶ should address this issue; work is under way along these lines. Compared to electrospinning²⁷, extrusion²⁴ and wet-spinning²⁵, this new approach to making fibrous structures is simple and scalable without the need for sophisticated tools.

Although polydispersity and molecular heterogeneity are often avoided in synthetic systems, here they facilitate equilibrium transitions among dramatically different morphological structures. This feature could be exploited to create soft materials—such as highly branched networks with uniform filament size—simply by tuning chain length distribution, temperature and surfactant concentration. Future studies of branching behaviours may provide new insights into drop assembly and stabilization, and, possibly, could reveal connections to molecular heterogeneity driven segregation and phase separation for function in biological matter^{28–30}. Moreover, the self-assembly processes are reversible, and the network structures can be permanently locked by ultraviolet crosslinking. The simple rules revealed by the experiments offer new concepts for creation of programmed spatio-temporal networks.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1809-8>.

1. Fasolo, M. & Sollich, P. Equilibrium phase behavior of polydisperse hard spheres. *Phys. Rev. Lett.* **91**, 068301 (2003).
2. Zaccarelli, E. et al. Crystallization of hard-sphere glasses. *Phys. Rev. Lett.* **103**, 135704 (2009).
3. Liddle, S. M., Narayanan, T. & Poon, W. C. K. Polydispersity effects in colloid-polymer mixtures. *J. Phys. Condens. Matter* **23**, 194116 (2011).
4. Lavrentovich, O. D. et al. Helical smectic A. *Europhys. Lett.* **13**, 313–318 (1990).

5. Gibaud, T. et al. Reconfigurable self-assembly through chiral control of interfacial tension. *Nature* **481**, 348–351 (2012).
6. Denkov, N., Tcholakova, S., Lesov, I., Cholakov, D. & Smoukov, S. K. Self-shaping of oil droplets via the formation of intermediate rotator phases upon cooling. *Nature* **528**, 392–395 (2015).
7. Guttman, S. et al. How faceted liquid droplets grow tails. *Proc. Natl Acad. Sci. USA* **113**, 493–496 (2016).
8. Toquer, G. et al. Colloidal shape controlled by molecular adsorption at liquid crystal interfaces. *J. Phys. Chem. B* **112**, 4157–4160 (2008).
9. Peddireddy, K., Kumar, P., Thutupalli, S., Herminghaus, S. & Bahr, C. Myelin structures formed by thermotropic smectic liquid crystals. *Langmuir* **29**, 15682–15688 (2013).
10. Ware, T. H., McConney, M. E., Wie, J. J., Tondiglia, V. P. & White, T. J. Voxellated liquid crystal elastomers. *Science* **347**, 982–984 (2015).
11. Gannon, M. G. J. & Faber, T. E. The surface tension of nematic liquid crystals. *Phil. Mag. A* **37**, 117–135 (1978).
12. Butt, H. J., Graf, K. & Kappl, M. *Physics and Chemistry of Interfaces* Ch. 3 (Wiley & Sons, 2006).
13. Schiele, K. & Trimper, S. On the elastic constants of a nematic liquid crystal. *Phys. Status Solidi B* **118**, 267–274 (1983).
14. Kamien, R. D. & Toner, J. Anomalous elasticity of polymer cholesterics. *Phys. Rev. Lett.* **74**, 3181–3184 (1995).
15. Kasten, H. & Strobl, G. Nematic wetting at the free surface of 4-cyano-4'-n-alkyl-biphenyls. *J. Chem. Phys.* **103**, 6768–6774 (1995).
16. Tintaru, M., Moldovan, R., Beica, T. & Frunza, S. Surface tension of some liquid crystals in the cyanobiphenyl series. *Liq. Cryst.* **28**, 793–797 (2001).
17. Lubensky, T. C., Petey, D., Currier, N. & Stark, H. Topological defects and interactions in nematic emulsions. *Phys. Rev. E* **57**, 610–625 (1998).
18. Allender, D. W., Crawford, G. P. & Doane, J. W. Determination of the liquid-crystal surface elastic constant K₂₄. *Phys. Rev. Lett.* **67**, 1442–1445 (1991).
19. Mušević, I. *Liquid Crystal Colloids* Ch. 7 (Springer, 2017).
20. Cholakov, D., Valkova, Z., Tcholakova, S., Denkov, N. & Smoukov, S. K. “Self-shaping” of multicomponent drops. *Langmuir* **33**, 5696–5706 (2017).
21. Shimshick, E. J. & McConnell, H. M. Lateral phase separation in phospholipid membranes. *Biochemistry* **12**, 2351–2360 (1973).
22. Sidky, H. & Whitmer, J. K. Elastic response and phase behavior in binary liquid crystal mixtures. *Soft Matter* **12**, 4489–4498 (2016).
23. Elias, F., Clarke, S. M., Peck, R. & Terentjev, E. M. Nematic order drives phase separation in polydisperse liquid crystalline polymers. *Macromolecules* **33**, 2060–2068 (2000).
24. Ahir, S. V., Tajbakhsh, A. R. & Terentjev, E. M. Self-assembled shape-memory fibers of triblock liquid-crystal polymers. *Adv. Funct. Mater.* **16**, 556–560 (2006).
25. Ohm, C. et al. Preparation of actuating fibres of oriented main-chain liquid crystalline elastomers by a wet spinning process. *Soft Matter* **7**, 3730–3734 (2011).
26. Xia, Y., Zhang, X. & Yang, S. Instant locking of molecular ordering in liquid crystal elastomers by oxygen-mediated thiol-acrylate click reactions. *Angew. Chem.* **130**, 5767–5770 (2018).
27. Krause, S., Dersch, R., Wendorff, J. H. & Finkelmann, H. Photocrosslinkable liquid crystal main-chain polymers: thin films and electrospinning. *Macromol. Rapid Commun.* **28**, 2062–2068 (2007).
28. Dufresne, E. R. et al. Self-assembly of amorphous biophotonic nanostructures by phase separation. *Soft Matter* **5**, 1792–1795 (2009).
29. Saranathan, V. et al. Structure, function, and self-assembly of single network gyroid (I4₃₂) photonic crystals in butterfly wing scales. *Proc. Natl Acad. Sci. USA* **107**, 11676–11681 (2010).
30. Radja, A., Horsley, E. M., Lavrentovich, M. O. & Sweeney, A. M. Pollen cell wall patterns form from modulated phases. *Cell* **176**, 856–868 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

Synthesis of NLCO drops

The NLCO emulsions were made in three steps: (1) preparation of emulsion drops containing monomer/chain-extender mixtures and suspended in an aqueous phase; (2) oligomerization within individual drops to link liquid crystal monomers together into liquid crystal oligomers (LCOs); and (3) creation of final NLCO drop suspensions in water, which are observed by video microscopy as a function of temperature.

Step 1

A 1:1 mixture of 1,4-bis-[4-(6-acryloyloxyhexyloxy)benzoyloxy]-2-methylbenzene (monomer RM82, Wilshire Technologies) and *n*-butylamine (chain-extender, Sigma-Aldrich) was dissolved in chloroform (Fisher Scientific); weight of chloroform is 3× the weight of RM82. A 0.2 wt% (of RM82 concentration) antioxidant, butylated hydroxytoluene (BHT, Sigma-Aldrich), was also added.

The resulting organic phase mixture was then made into a microdroplet emulsion in an aqueous solution. This was achieved by adding surfactant to the mixture of water and organic phase and then shaking to create polydisperse drops with diameters ranging from 10 to 100 μm . Three different surfactants were employed: negatively charged sodium dodecyl sulfate (anionic surfactant SDS, Sigma-Aldrich), positively charged hexadecyltrimethylammonium bromide (cationic surfactant C₁₆TAB, Sigma-Aldrich), and neutral polyoxyethylene (20) sorbitan monolaurate (Tween 20, Fisher Scientific). Although phenomena with the SDS surfactant are reported in the main text, experiments using the other two surfactants exhibited similar behaviour; spontaneous shape transitions are not restricted to a certain type of surfactant. However, in considering surfactants, it is important that homeotropic anchoring of NLCOs is favoured. All chemicals were used as received, without further purification/modification. At this stage, the suspended drops contain unchained monomers.

Step 2

The RM82/butylamine (chain-extender)/chloroform–SDS/water drop-let suspension is next placed in a water bath at 90 °C for oligomerization. During this process, the organic solvent (chloroform) evaporates completely from each emulsion drop, and then the butylamine begins to link the diacrylate RM82 molecules in chains via a self-catalysed ‘aza-Michael addition’ reaction. Afterwards, emulsion drops are composed of main-chain liquid crystal oligomers, LCOs. The degree of oligomerization, that is, the mean oligomer chain length, $\langle\ell\rangle$, is controlled by reaction time.

We note that the NLCO chain length distribution is broad and has a shorter mean oligomer chain length, $\langle\ell\rangle$, compared to previous liquid crystal elastomer work¹⁰.

Step 3

After the reaction, a small volume of the suspension (that is, the NLCO emulsion) is pipetted into a pre-heated glass well, sealed with coverslip and glue, and then cooled to 20 °C on a hotplate at a rate of $-1\text{ }^{\circ}\text{C min}^{-1}$. The suspension is then annealed (that is, quickly reheated to 90 °C and then slowly cooled to 20 °C again) before the drop morphology experiments are carried out.

Note, compared to pure RM82 monomer, which forms a nematic only above 86 °C (it is nematic in the 86–116 °C temperature range), the present LCOs exhibit a wide nematic window that extends below room temperature (that is, at least between 20–90 °C).

Drop morphology characterization

Sample cells containing NLCO drops were observed with a Leica DMIR13 inverted optical microscope in bright-field and polarization modes. Leica NPlan10×, Leica PL Fluotar L 63×, and Leica PL APO 100× objectives were used. Crossed-polarizers and full-wave retardation plates were deployed in the microscope to characterize director configurations within the drops.

Molecular weight measurements of NLCOs

To further characterize the LCO samples, most of the remaining emulsion solution was centrifuged (2,000 relative centrifugal force, RCF) at 15 °C and washed with deionized water several times to remove surfactant. It was then vacuum dried to remove water.

Part of the resultant bulk sample was then dissolved in tetrahydrofuran (THF; stabilized with 250 p.p.m. BHT, Alfa Aesar) at a concentration of 1.5 mg ml^{-1} for molecular weight analyses. The molecular weight analyses were performed using standard size exclusion chromatography (SEC, Tosoh Bioscience EcoSEC) with flow rate of 1.0 ml min^{-1} and with three analytical columns in series: TSKgel G2000HXL, TSKgel G2500HXL and TSKgel G3000HXL. A standard polystyrene sample (MW = 48,100 Da, Sigma-Aldrich) was also added during the SEC measurements for calibration purposes.

Macroscopic interfacial tension measurements

The interfacial tension arising between NLCO drop surfaces and background aqueous phase were measured macroscopically using the pendant drop method^{31,32}. The pendant drop technique works by analysing the shape of a large (millimetre-size) liquid drop hanging from a capillary tube (that is, a flat tip syringe needle), when it is about to detach from the capillary tube. The shape of the hanging drop depends on gravity, on cohesive forces of the drop within the background medium, and on the interfacial tension.

Our pendant drop tensiometry employs a custom-made temperature-regulated chamber, which enabled us to measure the system interfacial tension as a function of temperature. Each NLCO bulk sample was injected into a syringe coupled with a flat tip needle. The pendant NLCO drop was made in a 0.1 wt% SDS water solution in the temperature-regulated chamber. High contrast images of the drop contour were obtained while temperature was slowly decreased. The densities of the bulk NLCO were also measured at the different temperatures. Finally, using the methods and software created by Daerr et al.³², we fitted the drop contour for each sample at each temperature using the density of bulk NLCO and of the surrounding medium. With this information, the interfacial tension can be deduced (see measured interfacial tensions in Extended Data Fig. 3).

Macromer–monomer mixing experiments

These experiments were designed to confirm the influence of oligomer polydispersity on the shape transition phenomenology of NLCO drops. Specifically, we investigated how the average chain length of the oligomers in the NLCO mixture affects the shape transitions. For this study, however, we employed a synthetic approach to making the NLCO drop mixtures that was different from the methods in the main text.

We first synthesized a main-chain liquid crystal macromer with a number-average molecular weight of approximately 6,900 Da; following the work by Ware et al.¹⁰, this macromer had a mean chain length, $\langle\ell\rangle$, of ~ 9 , that is, a mean chain length roughly 9 times the monomer length. This sample had a polydispersity index (PDI) of 1.3 (for comparison with the NLCOs as described in the main text, see Extended Data Fig. 1). Note, PDI is defined as the ratio between weight-average molecular weight (\bar{M}_w) and number-average molecular weight (\bar{M}_n). $\text{PDI} = 1$ implies uniform polymer chain length.

The macromers thus obtained were then mixed with pure RM82 monomer at different weight ratios in chloroform (chloroform:liquid crystal mixture = 3:1 wt/wt). The mixture was then emulsified in 0.1 wt% SDS aqueous solution, and then the chloroform was evaporated in a 90 °C water bath. No further chain growth occurred during this process (confirmed by ¹H-NMR) because the chain-extender, *n*-butylamine, was not added to the mixture. When all chloroform was removed, temperature cycling was carried out following the same protocol described in step 3. The resulting drop shapes and micro-structures were observed in bright-field and polarization optical microscopy (POM).

Upon cooling, the macromer–monomer NLCO drops transitioned from smooth spheres to roughened spheres, flower-like structures, and filamentous structures (see Extended Data Fig. 4). These effects were dependent on the monomer:macromer weight ratio. Specifically, with increased mean oligomer chain length, $\langle \ell \rangle$, droplet structures with larger surface area are preferred (for example, longer, thinner filaments). The observed phenomenology agrees very well with the observations in the main text. We have argued that the increased bulk elasticity and decreased interfacial tension accompanying increased $\langle \ell \rangle$ promotes shape transitions. Quantitatively, as shown in Extended Data Fig. 4b, the droplet's tendency towards non-sphericity commences when $\langle \ell \rangle \approx 1.4$. This supplementary macromer–monomer mixing experiment thus exhibits the same tendency as the experiments that employed the synthesis-in-emulsion scheme. In the latter case, the droplet's tendency towards non-sphericity commenced when $\langle \ell \rangle \approx 1.5$. Note, the $\langle \ell \rangle$ at which non-sphericity commences can vary slightly with drop size.

We also carried out experiments with monomers only. Pure RM82 monomer ($\langle \ell \rangle = 1$) crystallizes during cooling; it is nematic only above 86 °C. Thus, shape transitions are not seen with drops containing only monomer. Furthermore, experiments with drops containing only macromers ($\langle \ell \rangle \approx 9$, see Extended Data Fig. 1) did not exhibit shape transitions. Evidently, the surface tension in drops with pure macromer is too large to permit shape transitions, presumably due to poor anchoring at the surface³³. Moreover, the macromer-only system viscosity is much larger²³ than the drops with smaller average chain length, making texture formation/reformation kinetics very slow.

By contrast, the NLCO mixtures we studied maintain good packing at the drop interface, probably because of different formation processes. The NLCOs were synthesized within each SDS stabilized drop. Thus, we expect that homeotropic director 'pre-alignment' could already exist at early stages of the oligomerization process in these drops, that is, when most of the oligomer chains are short and the surfactants can induce good anchoring. Thus, importantly, the monomers are approximately locked into their interfacial structure at the outset. Then as the chains grow longer throughout the drop, the longer oligomers spatially segregate to an already partially ordered surface at high coverage and can help increase order at the interface in a perturbative manner. This segregation and greater molecular packing/order reduces interfacial tension and ultimately induces shape transitions when interfacial tension becomes sufficiently small. In principle, the shape transformation process could still happen for a monodisperse system with very small surface tension. Our polydisperse NLCO mixtures, however, provide an easy segregation-driven route to lowering interfacial tension and elastic energy.

Calculation of system free energy

Here, we provide details of our model free energy calculations (see also the custom computer codes mentioned in the Code Availability section below). We employ equilibrium models because the effects were reversible with slow temperature cycling, and the structures were stable. Note, for rapid quenching (non-equilibrium) it is possible to generate a transient negative surface tension which could also drive the effect.

The Frank free energy for a nematic liquid crystal is given below; it accounts for the elastic energy associated with spatial distortions of the director \hat{n} in the liquid crystal³⁴, that is:

$$F_V = \frac{1}{2} \int d^3r \{ K_{11} (\hat{n} \cdot \nabla \hat{n})^2 + K_{22} (\hat{n} \cdot \nabla \times \hat{n})^2 + K_{33} [\hat{n} \times (\nabla \times \hat{n})]^2 - K_{24} \nabla \cdot [\hat{n} \times (\nabla \times \hat{n}) + \hat{n} (\nabla \cdot \hat{n})] \} \quad (1)$$

Here K_{11} , K_{22} and K_{33} are elastic constants for splay, twist and bend deformations, respectively. The final term, with the elastic constant K_{24} , is called the saddle-splay; it is absent from the corresponding Euler–Lagrange equation but contributes to the total free energy. Note, for simple solutions, the so-called splay-bend K_{13} elastic deformation and other second derivatives (or higher-order terms) of the director field

are usually not included—for example, inclusion of K_{13} without higher order terms can lead to paradoxes^{37,38}—nevertheless, interested readers can find free energy models which include K_{13} and discuss these issues³⁹. To simplify calculations even further, a one-constant limit of the Frank free energy¹⁷ is often applied, that is, K_{11} , K_{22} and K_{33} are set equal and expressed as a single value, K .

In addition to liquid crystal elastic free energy, our modelling includes interfacial free energy,

$$F_{\sigma-\text{iso}} = \gamma \int dS \quad (2)$$

and liquid crystal interface anchoring energy⁴⁰,

$$F_{\sigma-\text{aniso}} = \frac{1}{2} W_a \int dS \sin^2 \Phi \quad (3)$$

Here γ is the interfacial tension, W_a is the anchoring energy coefficient at the interface, and Φ is the angle between the liquid crystal director at the interface and the interface normal. Both terms are integrated over the surface.

Starting with these elastic energy expressions, and using well-established elastic models for a sphere (with radial director configuration)¹⁷ and for a cylindrical filament (with escaped-radial director configuration)^{18,39,41–43}, one can derive expressions for total system free energy.

For the sphere (one-constant approximation),

$$F_s = 8\pi \left(K - \frac{1}{2} K_{24} \right) R + \gamma 4\pi R^2 \quad (4)$$

Or, if $K_{11} \neq K_{33}$,

$$F_s = 8\pi \left(K_{11} - \frac{1}{2} K_{24} \right) R + \gamma 4\pi R^2 \quad (5)$$

For the filament (one-constant approximation),

$$F_f = \pi K \left(3 - \frac{K_{24}}{K} - \frac{1}{\sigma} \right) L + \gamma 2\pi r L \quad (6)$$

where $\sigma \equiv \frac{W_a r}{K} + \frac{K_{24}}{K} - 1 > 1$

Or, if $K_{11} < K_{33}$,

$$F_f = \pi K_{11} \left[2 + \frac{k}{\sqrt{k-1}} \tan^{-1} \sqrt{k-1} - \frac{k}{\sqrt{k-1}} \tan^{-1} \left(\frac{\sqrt{k-1}}{\sigma} \right) - \frac{K_{24}}{K_{11}} \right]$$

Or, if $K_{11} > K_{33}$,

$$F_f = \pi K_{11} \left[2 + \frac{k}{\sqrt{1-k}} \tanh^{-1} \sqrt{1-k} - \frac{k}{\sqrt{1-k}} \tanh^{-1} \left(\frac{\sqrt{1-k}}{\sigma} \right) - \frac{K_{24}}{K_{11}} \right] \quad (7)$$

Here,

$$k \equiv \frac{K_{33}}{K_{11}} \quad \text{and} \quad \sigma \equiv \frac{W_a r}{K_{11}} + \frac{K_{24}}{K_{11}} - 1 > 1$$

Above, K is the composite elastic constant (that is, in the one-constant approximation) of the NLCO mixture, R is the radius of the spherical NLCO drop, and r and L are the radius and length of the NLCO cylindrical filament, respectively. The twist elasticity, K_{22} , makes no contribution to the free energy of either the sphere or the cylinder, and the anchoring energy term (equation (3)) arises for cylinders with an escaped-radial director configuration, but not for spheres with radial director configurations. Note that in equations (1), (4) and (6), the saddle-splay (K_{24}) term is sometimes ignored, but for our system K_{24} is essential.

Ignoring saddle-splay would require a negative interfacial tension (γ) for the spontaneous shape transitions, which is inconsistent with our measurements of γ and with the reversible, equilibrium shape transition phenomenology observed.

Since we observed a spontaneous transition from spherical to filamentous drops during cooling, we anticipate that an energetically favourable spontaneous process must have $F_i < F_s$. We next examine the parameter values needed to favour this spontaneous process.

Among these parameters, γ was measured by the pendant drop technique using large (millimetre-size) drops. For example, the measured γ is 4.64 mN m⁻¹ at 80 °C and 2.36 mN m⁻¹ at 30 °C for the sample NLCO#2 which had 7 h of oligomerization time (see Methods and Extended Data Fig. 3; see refs.^{7,11,12,44,45} for details of this unusual temperature-dependent trend). In one (typical) sample, R and r were measured by bright-field optical microscopy to be 13.0 μ m and 0.65 μ m, respectively. Because it is difficult to directly measure the elastic constants and anchoring energy coefficients of our NLCOs, we estimate their values at room temperature: $K \approx 10^{-10}$ N and $W_a \approx 10^{-4}$ J m⁻². These numbers are chosen based on the values, and on the relationships between parameters, in drops of monodisperse small molecule liquid crystals. When interfacial free energy is dominant (for example, at high temperature or with short $\langle \ell \rangle$), we expect the NLCO droplets to remain spherical. Then, knowing the interfacial tension, we can compute an estimate for K . Similarly, since the droplet adopts a radial configuration with a hedgehog defect at its centre¹⁹, W_a is estimated by assuming that the extrapolation length, $\xi_s \equiv K/W_a \approx 1$ μ m, for the spherical drop is much smaller than the drop diameter. Note that W_a typically ranges from 10^{-6} to 10^{-3} J m⁻² for weak to strong anchoring. (For reader reference, we list experimental values for important parameters. For thermotropic liquid crystals: (1) 5CB has $K_{11} \approx 6.2 \times 10^{-12}$ N and $K_{33} \approx 8.25 \times 10^{-12}$ N (-25 °C); (2) 8CB has $K_{11} \approx 6.0 \times 10^{-12}$ N and $K_{33} \approx 6.4 \times 10^{-12}$ N (-35 °C)⁴⁶. For lyotropic liquid crystals: (1) DSCG has $K_{11} \approx 10.2 \times 10^{-12}$ N and $K_{33} \approx 24.9 \times 10^{-12}$ N (16 wt%, within -4 °C of T_{NI}); (2) SSY has $K_{11} \approx 4.3 \times 10^{-12}$ N and $K_{33} \approx 6.1 \times 10^{-12}$ N (29 wt%, within -2 °C of T_{NI})⁴⁷. For liquid crystal monomers that are molecularly similar to RM82: $K \approx 2.6 \times 10^{-12}$ N (-97 °C)⁴⁸. For anchoring energy: (1) 5CB/SiO interfaces have $W_a \approx 4.0 \times 10^{-5}$ J m⁻² (-35 °C)⁴⁹; liquid crystal mixture (Merck-BDH, MLC 6608)/lecithin interfaces have $W_a \approx 4.6 \times 10^{-4}$ J m⁻² (23.1 °C)⁵⁰; 8OCB/DMOAP-treated-glass interfaces have $W_a \approx 1.0 \times 10^{-4}$ J m⁻² (-75 °C)⁵¹.)

The models and criteria above, along with the experimental and estimated parameters, lead to well-defined constraints. A spontaneous shape transition is only possible, for example, when $K_{24} \geq 32 \times K_{11}$, which is far too large for most liquid crystal materials. Note that abandoning the one-constant approximation does not affect this conclusion. In addition, varying K and W_a over a reasonable range—for example, from 5×10^{-11} to 10^{-10} N and 5×10^{-5} to 5×10^{-4} J m⁻², respectively—does not affect the conclusion (see details in Extended Data Table 1). Revisiting the initial comparison between equations (4) and (6) suggests that a smaller ‘true’ γ is needed to greatly increase the probability of a shape transition. We next consider how this situation can be realized through oligomer polydispersity and spatial segregation.

In applying equations (4) and (6), our models treated the bulk NLCO as a homogeneous material with monodisperse chain length. However, the NLCO mixture is polydisperse; it is composed of oligomers with a broad distribution of chain lengths. This new degree of freedom offers the possibility for spatial rearrangement of the oligomers within the drops, wherein the long-chain-length oligomers move preferentially to the interface and the short-chain-length oligomers move to be closer to the drop centre. For micrometre-size NLCO emulsion drops, molecules can easily diffuse and segregate within the confining structure on experimental timescales.

On the basis of simple energy considerations, spontaneous segregation to produce a long-chain-rich shell (near the surface or interface) and a short-chain-rich core (for example, near the sphere centre, or central

axis of a filament or cylinder) will lower the system free energy. Consider a quantitative example. For spherical drops of radius R , one can compute the elastic free energy density in the shell and core regions using the relation: $F_{\text{elastic}} = 8\pi K_{11}R$. If we define the core as the central region with a radius ranging from 0 to $R/2$, and the shell as the remainder of the sphere, then the splay elastic energy density in the core is about 7 \times greater than that in the shell. In our experiments, the bulk elastic energy is lowered because short-chain NLCOs, which have smaller elastic constants, preferentially occupy the core regions which have very substantial director distortions (that is, large splay or splay/bend elastic distortions). This phenomenon can be understood as a chain-length-dependent ‘driving force’ due to the elastic energy density gradient within the drop. This energy gradient between core and shell regions leads to spatial rearrangement of oligomers with different chain lengths.

Moreover, the segregation of long-chain oligomers to the drop surface will reduce the interfacial energy of the drop (compared to a drop containing a homogeneous mixture, as is the case for the millimetre-size drops). We thus expect $\gamma_{\text{true;micro-emulsion}}$ to be less than the corresponding millimetre-size droplet interfacial tension, $\gamma_{\text{measured;mm-drop}}$, measured in the pendant drop experiments: $\gamma_{\text{true;micro-emulsion}} < \gamma_{\text{measured;mm-drop}}$. NLCO polydispersity, and its consequences for chain-length-dependent spatial segregation of oligomers within the drop, generates a critical new feature in the microemulsion that influences shape transitions and self-assembly. The resultant reduction of interfacial tension reduces the unphysically large K_{24} requirement (calculated above) and thus resolves concerns raised by the simple shape instability model calculations for homogeneous, monodisperse liquid crystals. For example, using the same parameters for K and W_a , but with a reduction of γ by 10 \times , we obtain the requirement $K_{24} \geq 5.5 \times K_{11}$, which is in an acceptable range for saddle-splay modulus. Of course, further reduction of γ (as long as interfacial energy still dominates at high temperature) decreases the required K_{24} value even more and renders shape transitions to be even more likely. (Note also that the segregation-induced decrease in bulk elastic energy will further reduce the requirements for γ reduction, because the decrease in elastic free energy in filaments is greater than in spherical drops.)

We can estimate the reduction of γ induced by oligomer segregation using a simple bidisperse demixing model (alluded to elsewhere in Methods). In the model, one component of the mixture is a monomer ($\langle \ell \rangle = 1$), and the other component is a macromer with $\langle \ell \rangle \approx 9$. As shown in Extended Data Fig. 4c, for a monomer:macromer mixture weight ratio of 1:0.7, the overall mean chain length is $\langle \ell \rangle_{\text{whole}} \approx 1.53$. This condition gives rise to filamentous drop structures. Taking this condition to be exemplary for filament formation, we next show how the macromers and monomers segregate into core and shell regions with uneven chain length distributions in order to lower overall system elastic energy.

Because the total amount of monomer and macromer is conserved during segregation, we readily generate an equation relating the monomer:macromer weight ratio and $\langle \ell \rangle_{\text{whole}}$ of the initial homogeneously distributed mixture to the final, segregated long-chain-rich shell with $\langle \ell \rangle_{\text{shell}}$ and the short-chain-rich core with $\langle \ell \rangle_{\text{core}}$. The resulting relation is:

$$\frac{0.7}{1+0.7} = \frac{V_{\text{shell}}}{V_{\text{total}}} \times \frac{6,641(\langle \ell \rangle_{\text{shell}} - 1)}{6,641(\langle \ell \rangle_{\text{shell}} - 1) + 673(9 - \langle \ell \rangle_{\text{shell}})} + \frac{V_{\text{core}}}{V_{\text{total}}} \times \frac{6,641n}{6,641n + 673(1 - n)} \quad (8)$$

Here n is the number fraction of macromer in the core region, and V_{shell} and V_{core} are volume of shell and core regions, respectively: $V_{\text{shell}} + V_{\text{core}} = V_{\text{total}}$. The mean chain length of the core can also be written in a simple form:

$$\langle \ell \rangle_{\text{core}} = 1 \times (1 - n) + 9n \quad (9)$$

For simplicity, we assume that segregation leads to equal volumes of core and shell, and we set $\langle \ell \rangle_{\text{shell}} = 1.70$. Equations (8) and (9) then give $n \approx 4.9\%$ and $\langle \ell \rangle_{\text{core}} \approx 1.39$. Note that it is straightforward to use this model with other assumptions about the volume segregation ratio and the mean oligomer chain length in the shell; the general qualitative conclusions about system free energy reduction (compared to the homogeneous drop) will be the same for sensible parameters.

To quantify how the chain length redistribution within the structure affects the system's overall bulk elastic energy, we take $\langle \ell \rangle_{\text{core}} \approx 1.39$ and $\langle \ell \rangle_{\text{shell}} = 1.70$, and we compute the elastic free energy volume integrals over the target geometry (for example, radial/sphere and escaped-radial/filament) using the Frank free energy equation (equation (1)). We then compare the Frank free energy of the segregated system to that of the homogeneously distributed system with $\langle \ell \rangle_{\text{whole}} \approx 1.53$. For this calculation, we assume that the elastic constant is proportional to oligomer mean chain length^{14,52}. The comparison reveals that this reduction in $\langle \ell \rangle_{\text{core}}$ will decrease bulk elastic energy by $\sim 5\%$ for the sphere and $>8\%$ for the filament (that is, compared to the homogeneously distributed systems).

Most importantly, as a result of oligomer segregation (for example, an increase of local mean chain length near the interface from 1.53 to 1.70, by ~ 0.17), the interfacial tension is lowered dramatically. To better appreciate this assertion, consider the homogeneous mixtures NLCO#2 (7 h oligomerization) and NLCO#5 (24 h oligomerization) in Extended Data Fig. 3. For these samples, the interfacial tension (for example, at 30°C) of the drops made with NLCO#5 (0.12 mN m^{-1}) is much smaller, by a factor of $\sim 20\times$, compared to that of NLCO#2 (2.36 mN m^{-1}), whereas the average chain length changed by only ~ 0.10 . Indeed, from the pendant drop experiments, one would expect that a long-chain-rich interface (with local increase of mean chain length of ~ 0.17) would reduce γ substantially, that is, γ would be reduced by at least a factor of $20\times$. Hence, our earlier statement in the text that a “conservative reduction of γ by ten times to approximately 0.1 mN m^{-1} yields sphere–filament instabilities with reasonable saddle-splay moduli” is a conservative estimate of the expected change of interfacial tension.

Last, we reiterate that in the millimetre-size drops measured by the pendant technique, we expect the chain-length-dependent segregation to be insignificant. Macroscopically, the drop is approximately homogeneous because the millimetre-size drops have many nematic microdomains (each with a director oriented uniformly along a different direction) and many randomly situated disclination lines. The larger drops have lower curvatures too, and therefore director distortion is reduced near their surfaces. The largest distortions that could drive segregation should occur locally near disclinations and at domain walls distributed roughly uniformly throughout the drop; thus, they will not promote net migration to the surface. Moreover, the required molecular diffusion over long distances will be small because of extremely long diffusion times (domain walls would also act as barriers to oligomer diffusion, and so on). Thus, interfacial tension derived from the pendant drop experiments is set by the macroscopic ‘average’ of the whole droplet, that is, $\langle \ell \rangle$ of the initial (source) NLCO distributions.

Crosslinking NLCO structures into NLCEs

To crosslink structural NLCOs into solid NLCE structures, 2 wt% (by RM82 concentration) of 2,2-dimethoxy-2-phenylacetophenone (radical photoinitiator, Sigma-Aldrich) was added into the initial RM82/butylamine/chloroform mixture. Then the same oligomerization-cooling method described above was performed. After the NLCO emulsions reached their final (desired) equilibrium morphologies, the NLCOs were radically crosslinked into elastomers by UV radiation (365 nm) for a few minutes. We then evaporated the background solution, sputter-coated thin metallic films (gold or iridium) onto the structures, and observed the resulting NLCEs via SEM (dual-beam FEI Strata DB235 Focused Ion Beam/SEM) using a 5.00 kV electron beam.

Data availability

The authors declare that the data supporting the findings of this study are available within the text, including the Methods section, and Extended Data files. Raw data are available from the corresponding author upon reasonable request.

Code availability

Custom computer codes associated with modelling in this study are available on GitHub (<https://github.com/wei-shao-wei/Molecular-heterogeneity-induces-reconfigurable-nematic-liquid-crystal-drops>).

- Adamson, A. W. & Gast, A. P. *Physical Chemistry of Surfaces* 6th edn, Ch. 2 (Wiley & Sons, 1997).
- Daerr, A. & Mogne, A. Pendant_drop: an imagej plugin to measure the surface tension from an image of a pendant drop. *J. Open Res. Softw.* **4**, e3 (2016).
- Terentjev, E. M. Density functional model of anchoring energy at a liquid crystalline polymer-solid interface. *J. Phys. II France* **5**, 159–170 (1995).
- Frank, F. C. I. Liquid crystals: on the theory of liquid crystals. *Discuss. Faraday Soc.* **25**, 19–28 (1958).
- Oseen, C. W. The theory of liquid crystals. *Trans. Faraday Soc.* **29**, 883–899 (1933).
- Nehring, J. & Saupe, A. On the elastic theory of uniaxial liquid crystals. *J. Chem. Phys.* **54**, 337–343 (1971).
- Barbero, G. & Oldano, C. Derivative-dependent surface-energy terms in nematic liquid crystals. *Nuovo Cimento D* **6**, 479–493 (1985).
- Sparavigna, A., Komitov, L. & Strigazzi, A. Hybrid aligned nematics and second order elasticity. *Phys. Scr.* **43**, 210 (1991).
- Crawford, G. P., Allender, D. W. & Doane, J. W. Surface elastic and molecular-anchoring properties of nematic liquid crystals confined to cylindrical cavities. *Phys. Rev. A* **45**, 8693–8708 (1992).
- Rapini, A. & Papoular, M. Distorsion d'une lamelle nématique sous champ magnétique conditions d'ancrage aux parois. *J. Phys. Colloques* **30**, 54–56 (1969).
- Cladis, P. E. & Kléman, M. Non-singular disclinations of strength $S = +1$ in nematics. *J. Phys. France* **33**, 591–598 (1972).
- Meyer, R. B. On the existence of even indexed disclinations in nematic liquid crystals. *Phil. Mag.* **27**, 405–424 (1973).
- Burylov, S. V. Equilibrium configuration of a nematic liquid crystal confined to a cylindrical cavity. *J. Exp. Theor. Phys.* **85**, 873–886 (1997).
- Li, X. & Denn, M. M. Interface between a liquid crystalline polymer and a flexible polymer. *Macromolecules* **35**, 6446–6454 (2002).
- Wu, J. & Mather, P. T. Interfacial tension of a liquid crystalline polymer in an isotropic polymer matrix. *Macromolecules* **38**, 7343–7351 (2005).
- Bradshaw, M. J., Raynes, E. P., Bunning, J. D. & Faber, T. E. The Frank constants of some nematic liquid crystals. *J. Phys. France* **46**, 1513–1520 (1985).
- Zhou, S. et al. Elasticity, viscosity, and orientational fluctuations of a lyotropic chromonic nematic liquid crystal disodium cromoglycate. *Soft Matter* **10**, 6571–6581 (2014).
- Xia, Y., Serra, F., Kamien, R. D., Stebe, K. J. & Yang, S. Direct mapping of local director field of nematic liquid crystals at the nanoscale. *Proc. Natl Acad. Sci. USA* **112**, 15291–15296 (2015).
- Yokoyama, H. & van Sprang, H. A. A novel method for determining the anchoring energy function at a nematic liquid crystal-wall interface from director distortions at high fields. *J. Appl. Phys.* **57**, 4520–4526 (1985).
- Yang, F., Ruan, L. & Sambles, J. R. Homeotropic polar anchoring energy of a nematic liquid crystal using the fully leaky waveguide technique. *J. Appl. Phys.* **88**, 6175–6182 (2000).
- Škarabot, M., Osmanagić, E. & Mušević, I. Surface anchoring of nematic liquid crystal 8OCB on a DMOAP-silanated glass surface. *Liq. Cryst.* **33**, 581–585 (2006).
- Ciferri, A., Krigbaum, W. R. & Meyer, R. B. *Polymer Liquid Crystals* Ch. 6 (Academic, 1982).

Acknowledgements We thank the following for discussions: K. B. Aptowicz, C.-C. Chang, P. J. Collings, A. de la Cotte, Z. S. Davidson, R. Dreyfus, P. Habdas, A. Hill, Y.-Y. Ho, R. D. Kamien, J. Jeong, T. C. Lubensky, X. Ma, A. Martinez, C. K. Mishra and P. Palffy-Muhoray. A. Soleymannezhad and J. Timmons (Tosoh Bioscience) assisted with SEC operation and analysis. We acknowledge financial support from the National Science Foundation (grant no. DMR16-07378), the Materials Research Science & Engineering Center (MRSEC) at University of Pennsylvania (grant no. DMR-1720530) including MRSEC's Optical Microscopy and Electron Microscopy Shared Experimental Facility, and NASA (grant no. 80NSSC19K0348).

Author contributions W.-S.W., Y.X., S.Y. and A.G.Y. conceived the idea and designed the experiments. W.-S.W., Y.X. and S.E. initiated and performed the experiments. W.-S.W., Y.X., S.E., S.Y. and A.G.Y. worked on different facets of the data analysis. W.-S.W. and A.G.Y. wrote the paper, and all authors contributed to the final manuscript.

Competing interests The authors declare no competing interests.

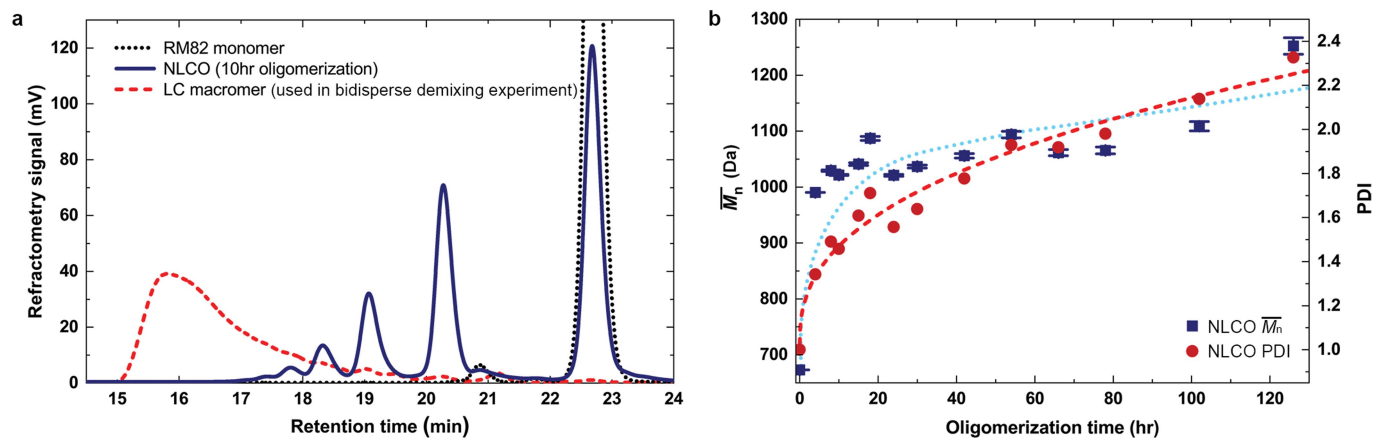
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1809-8>.

Correspondence and requests for materials should be addressed to W.-S.W.

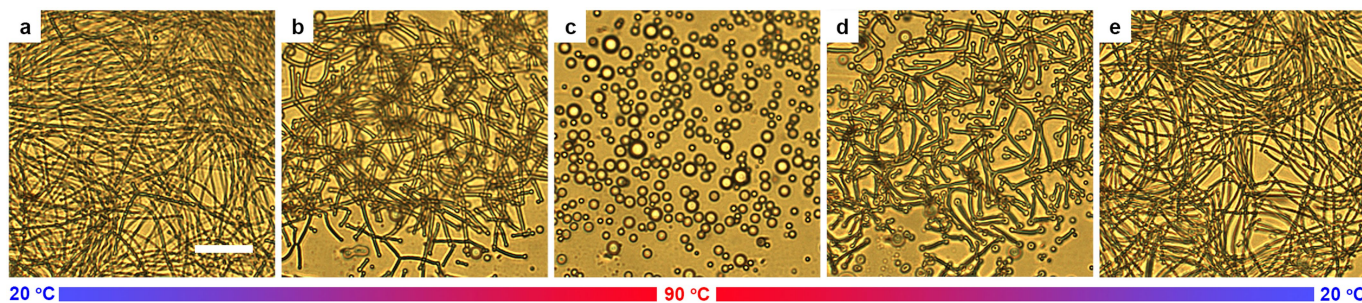
Peer review information Nature thanks Kari Dalnoki-Veress, Stoyan Smoukov and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



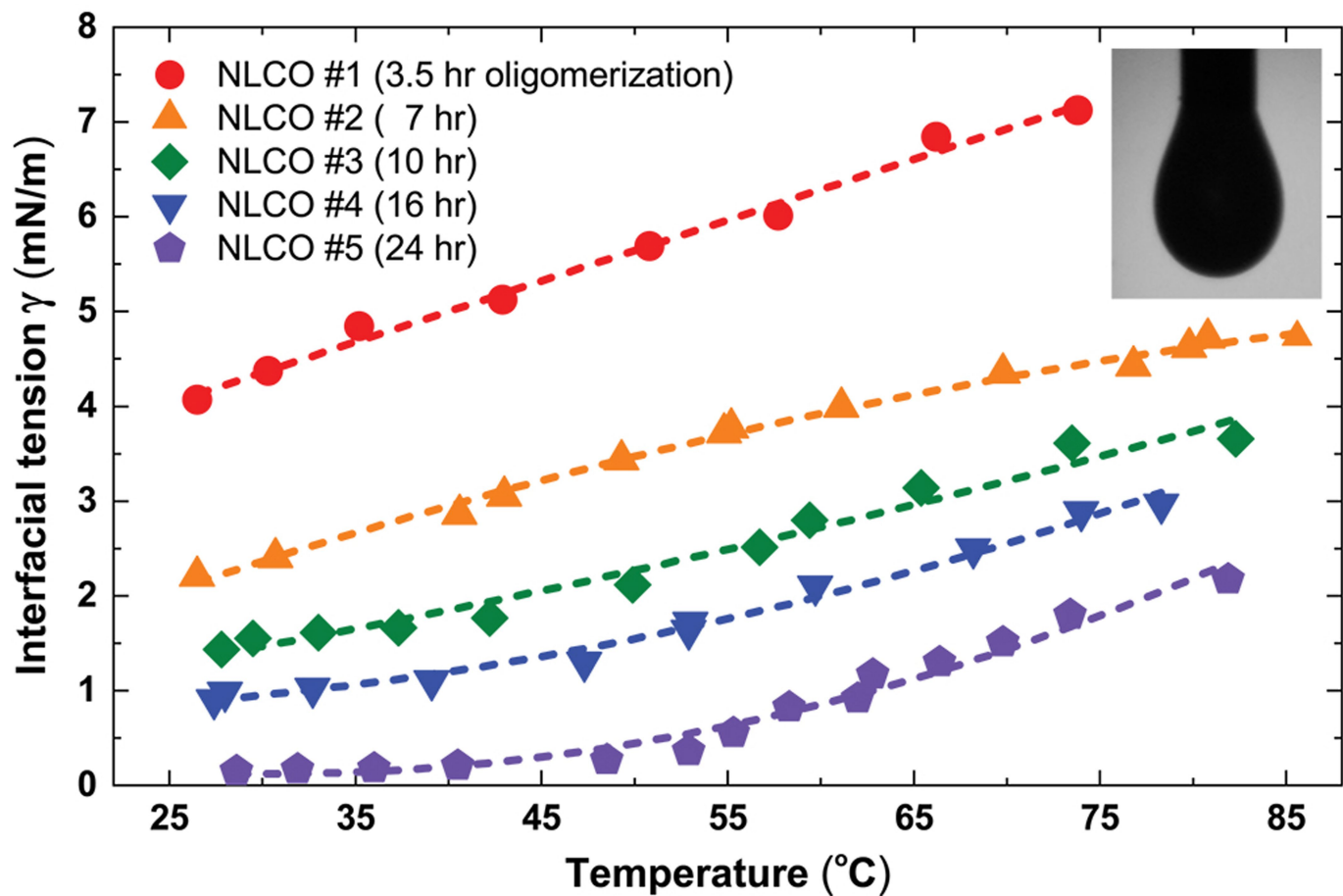
Extended Data Fig. 1 | Distribution of chain length and molecular weight of NLCO samples measured by size exclusion chromatography. **a**, The NLCO source samples are shown by size exclusion chromatography (SEC) to be a mixture of monomers, dimers, trimers, tetramers and other oligomers (blue solid line; peaks appearing at longer retention times represent shorter chain lengths). In step-polymerization processes, when the extent of reaction is less than 0.9, monomers dominate the overall molar fraction. For comparison, liquid crystal (LC) macromers (used in the supporting experiment in Methods) synthesized following the schemes of Ware et al.¹⁰ are also included in the plot; these show longer mean chain length and very few (if any) short-chain components (red dashed line). Furthermore, since our system is made in an

aqueous solution, the polymerization rate is expected to be slower. For reference, the black dotted line shows the peak for pure RM82 monomer. **b**, Calculated from SEC data (example in **a**), the number-average molecular weight (\bar{M}_n , blue solid squares) and polydispersity index (PDI, red filled circles) of the NLCOs are shown as function of oligomerization time. Both \bar{M}_n and PDI increase with oligomerization time. The dotted (for \bar{M}_n) and dashed (for PDI) curves are to guide the eye. (For comparison, the LC macromer synthesized following the schemes of Ware et al.¹⁰ and used in Methods section ‘Macromer–monomer mixing experiments’ has $\bar{M}_n \approx 6,900$ Da and PDI ≈ 1.3 .) Bars indicate the spread in \bar{M}_n , which mainly arises from our inability to detect the longer-chain components.



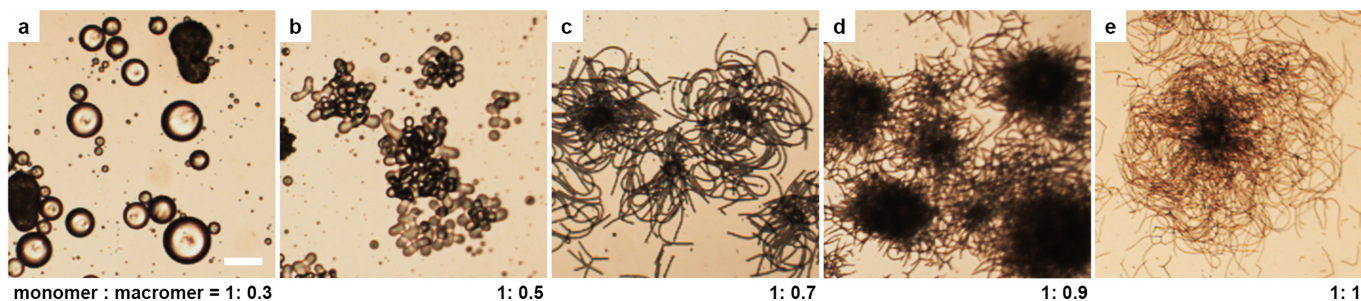
Extended Data Fig. 2 | Bright-field optical microscopy images showing reversible shape transitions of NLCO drops during temperature cycling. **a–c**, When temperature is increased from room temperature (20 °C) to a higher value (here 90 °C), the NLCO filamentous structures reversibly evolve into spherical microdroplets. **c–e**, When cooled from 90 °C to 20 °C, the spherical

microdroplets reversibly evolve back into filamentous structures. The drop morphology can be transformed repeatedly, remaining quantitatively similar. Here, multiple small drops evolve in the field of view; data shown earlier (Fig. 1c–h) showed only one large evolving drop. Scale bar in **a** (for all panels), 20 μm .



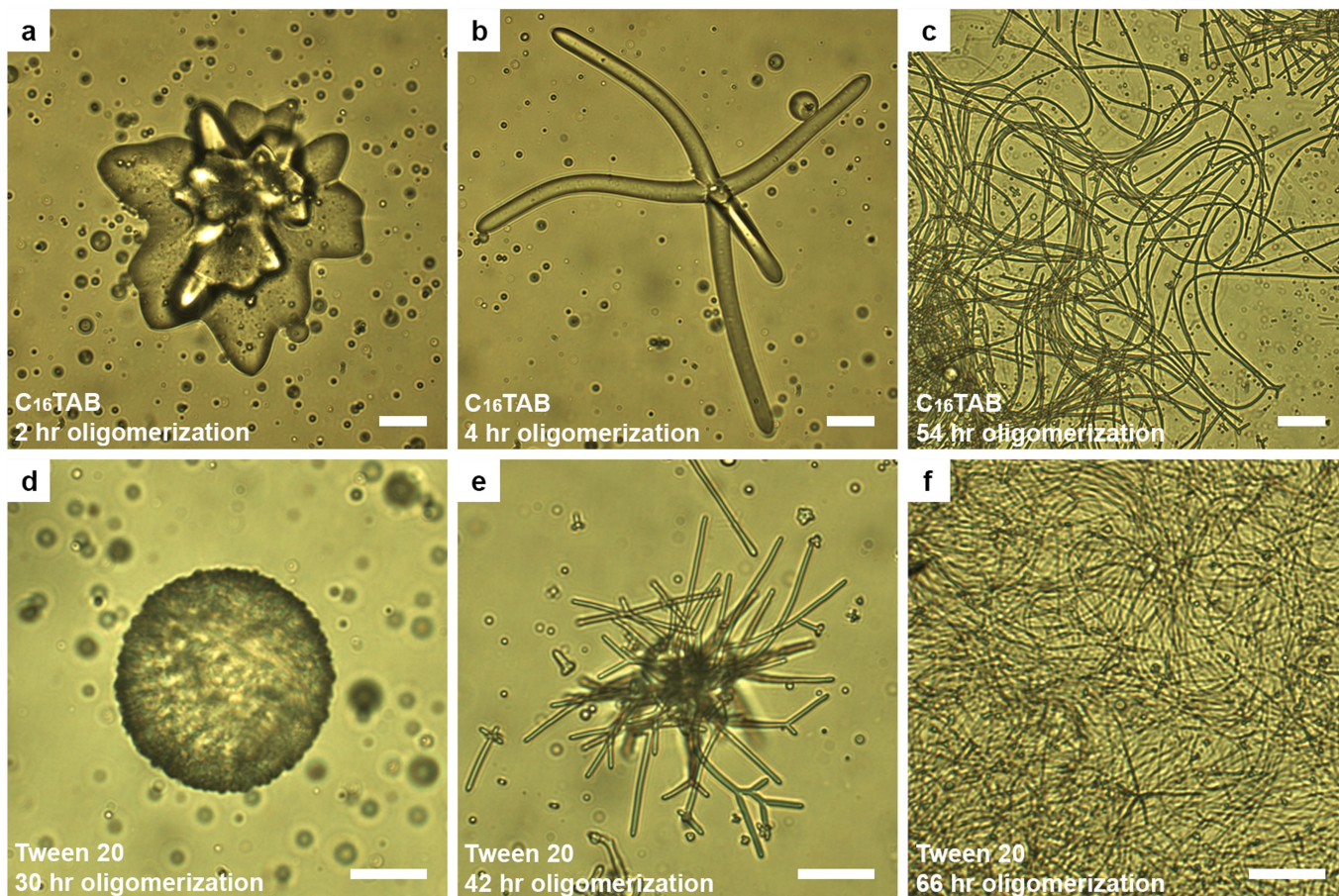
Extended Data Fig. 3 | Macroscopic interfacial tension of NLCO pendant drops as function of temperature and NLCO oligomerization time. The NLCO drop has homeotropic anchoring at the interface in an aqueous solution consisting of 0.1 wt% SDS. The interfacial tension γ decreases with decreasing temperature and increasing NLCO mean oligomer chain length $\langle \ell \rangle$ (consult

Fig. 2f for the relation between oligomerization processing time and $\langle \ell \rangle$). Inset, optical image of a NLCO pendant drop hanging from a flat-tip syringe needle (1.26 mm outer diameter) in a 0.1 wt% SDS aqueous solution. The dashed curves are to guide the eye.



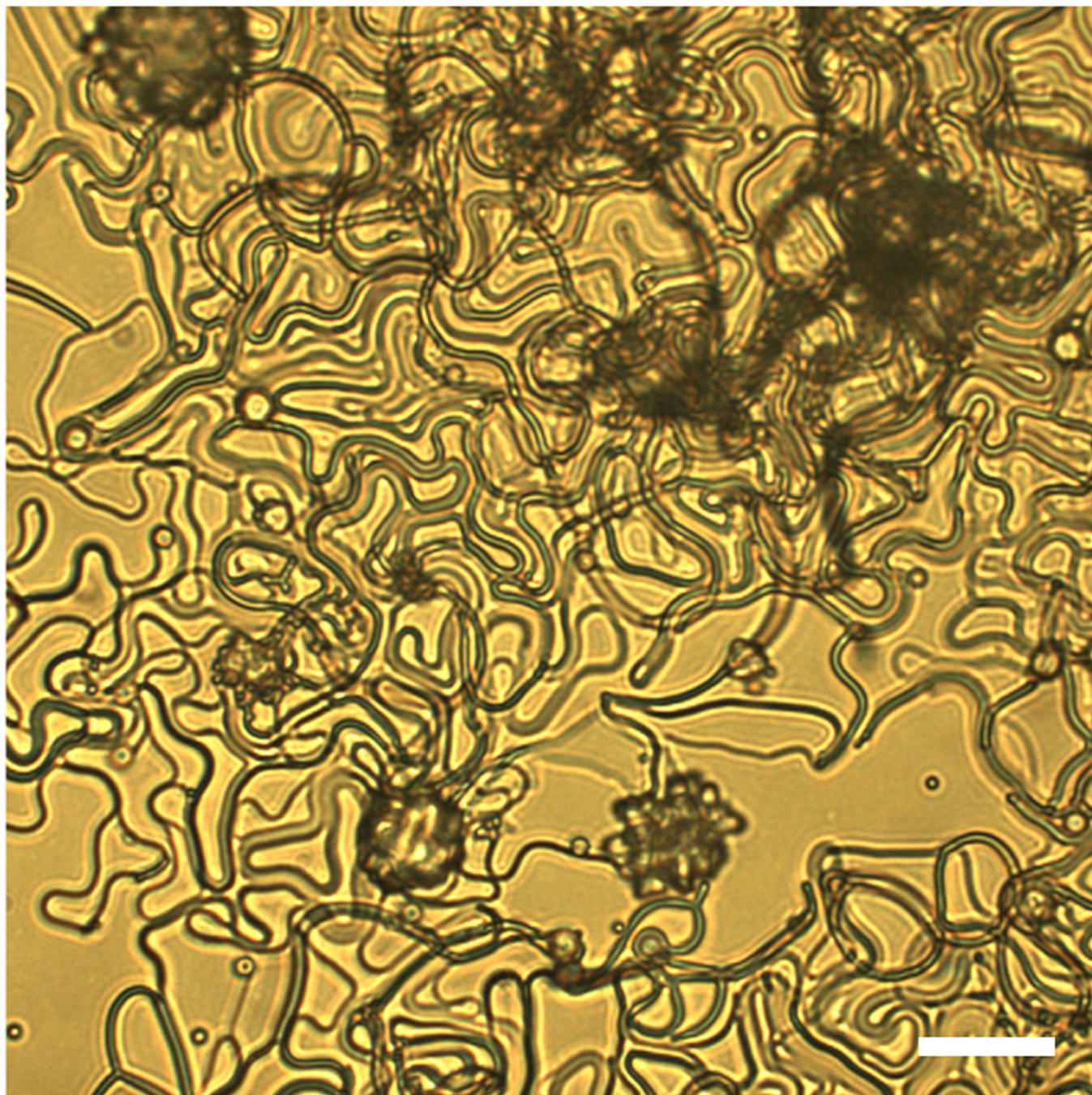
Extended Data Fig. 4 | Bright-field optical microscopy images of drop morphologies obtained from mixtures of macromers ($\langle \ell \rangle \approx 9$) and monomers (RM82) at different weight ratios in a 0.1 wt% SDS aqueous solution after cooling. a–e, Images for monomer:macromer weight ratios of 1:0.3, 1:0.5, 1:0.7, 1:0.9 and 1:1, respectively. With a fixed amount of RM82, increasing the

concentration of macromers in the drop leads to longer $\langle \ell \rangle$, larger bulk elasticity and lower interfacial tension. The last two factors favour interfacial roughening and filament formation. Images are taken at room temperature after cooling. Scale bar in **a** (for all panels), 50 μm .



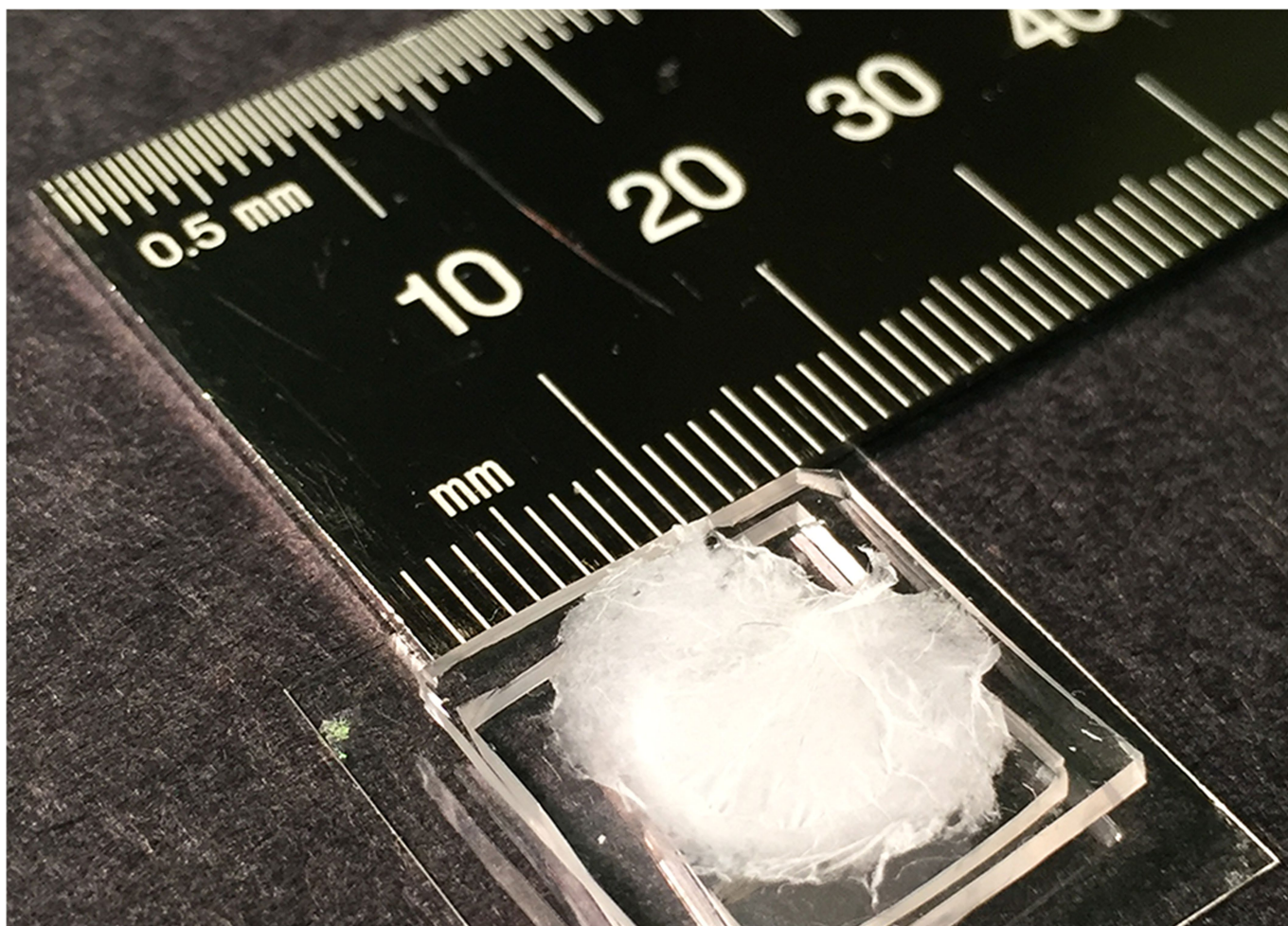
Extended Data Fig. 5 | Bright-field optical microscopy images of NLCO structures in aqueous solutions of different surfactants as a function of the mean oligomer chain length, $\langle \ell \rangle$. a–f, NLCO drops in aqueous solutions with different surfactants, that is, with either cationic (a–c; hexadecyltrimethylammonium bromide, $C_{16}TAB$) or nonionic (d–f, Polysorbate 20, Tween 20) surfactants. These systems exhibit a drop morphology evolution similar to that resulting from SDS, that is, with respect to cooling and an increase of $\langle \ell \rangle$. After cooling from 90 °C to 20 °C, a representative NLCO drop in

a 0.5 mM $C_{16}TAB$ aqueous solution (below the CMC; a–c) and a representative NLCO drop in a 0.03 mM Tween 20 aqueous solution (below the CMC; d–f). Both evolve with increasing $\langle \ell \rangle$ (left to right; consult Fig. 2f for the relation between the oligomerization time (shown at bottom left in all panels) and $\langle \ell \rangle$). For this behaviour to be shown, it is important that NLCOs should favour homeotropic anchoring at the drop interface. Images are taken at room temperature. Scale bars in a–f, 20 μm .



Extended Data Fig. 6 | Bright-field optical microscopy image of aggregated NLCO filamentous structures for an SDS concentration above the CMC. In a 1 wt% SDS aqueous solution (shown here), the NLCO drops exhibit a similar but more complicated shape transition behaviour than that which occurs below the CMC, after cooling from 90 °C to 20 °C. For example, aggregated

filamentous structures form and sometimes stick to the substrate due, in part, to micelle-induced depletion. (For comparison, expanded filamentous structures form in aqueous solutions below the CMC, see example in Fig. 1h.) Scale bar, 20 μm .



Extended Data Fig. 7 | Self-assembled NLCE fibrous mat. NLCE fibres can be densely packed into centimetre-wide and few-micrometres-thick, non-woven, free-standing mats by sedimentation. Here we show an image of such a mat

(mounted on a hollowed holder) with a diameter greater than 1 cm. Corresponding high-magnification SEM images of this mat are shown in Fig. 4e, f.

Extended Data Table 1 | Parameter values that permit shape transitions of NLCO drops

$\gamma_{\text{measured; mm-drop}}$ (mN/m)	K (N)	W_a (J/m ²)	Required K_{24} for shape transition	Reduction factor of $\gamma_{\text{measured; mm-drop}}$ gives $\gamma_{\text{true; micro-emulsion}}$	Required K_{24} for shape transition with reduced γ
2.36	5×10^{-11}	$5 \times 10^{-5} \sim 5 \times 10^{-4}$	$60.2 \times K_{11}$	20x	$5.7 \times K_{11}$
2.36	10^{-10}	$5 \times 10^{-5} \sim 5 \times 10^{-4}$	$31.6 \times K_{11}$	10x	$5.7 \times K_{11}$
2.36	5×10^{-10}	5×10^{-4}	$8.6 \times K_{11}$	2x	$5.7 \times K_{11}$

This table summarizes parameter values that permit shape transitions (with reasonable saddle-splay elastic constants, K_{24}) based on our model free energy calculations. The calculations fix the interfacial tension ($\gamma_{\text{measured; mm-drop}}$; column 1) at 30 °C measured by the pendant drop technique. A fairly wide range of estimated elastic constants (K ; column 2) and anchoring energy coefficients (W_a ; column 3) are employed in the calculations. In line with the criteria for a spontaneous shape transition, we require the required saddle-splay elastic constant (K_{24} ; column 4) to be of the order of ~6 times the splay modulus K_{11} (or less). The calculations show that a smaller ‘true’ interfacial tension, $\gamma_{\text{true; micro-emulsion}}$ (expressed as a reduction factor of $\gamma_{\text{measured; mm-drop}}$; column 5), will relax the saddle-splay requirement (column 6). This reduction of γ can be realized through oligomer polydispersity and the resultant oligomer spatial segregation in the elastic stress field. Note that a reasonable range for K is 5×10^{-11} N to 10^{-10} N; a value of $K = 10^{-11}$ N is probably too small, as it is the same order of magnitude as small molecule liquid crystals such as 5CB.

Probing the critical nucleus size for ice formation with graphene oxide nanosheets

<https://doi.org/10.1038/s41586-019-1827-6>

Guoying Bai^{1,2}, Dong Gao³, Zhang Liu¹, Xin Zhou^{4,5,6*} & Jianjun Wang^{1,6,7*}

Received: 2 October 2018

Accepted: 17 September 2019

Published online: 18 December 2019

Water freezing is ubiquitous and affects areas as diverse as climate, the chemical industry, cryobiology and materials science. Ice nucleation is the controlling step in water freezing^{1–5} and has, for nearly a century, been assumed to require the formation of a critical ice nucleus^{6–10}. But there has been no direct experimental evidence for the existence of such a nucleus, owing to its transient and nanoscale nature^{6,7}. Here we report ice nucleation in water droplets containing graphene oxide nanosheets of controlled sizes and show that they have a notable impact on ice nucleation only above a certain size that varies with the degree of supercooling of the droplets. We infer from our experimental data and theoretical calculations that the critical size of the graphene oxide reflects the size of the critical ice nucleus, which in the case of sufficiently large graphene oxides sits on their surface and gives rise to ice formation behaviour consistent with classical nucleation theory. By contrast, when the graphene oxide size is smaller than that of the critical ice nucleus, pinning at the periphery of the graphene oxide deforms the ice nucleus as it grows. This gives rise to a much higher free-energy barrier for nucleation and suppresses the promoting effect of the graphene oxide¹¹. The results provide experimental information on the existence and temperature-dependent size of the critical ice nucleus, which has previously only been explored theoretically and through simulations^{12–16}. As pinning of a pre-critical nucleus at a nanoparticle edge is not specific to the ice nucleus on graphene oxides, we expect that our approach could be extended to probe the critical nuclei in other nucleation processes.

Theory¹⁷ and experiment¹⁸ have shown that for radii ranging from around 10 Å to 1,000 Å, size profoundly influences a particle's ability to induce ice nucleation. Such a size effect is evident when we consider that antifreeze proteins (AFPs) suppress ice formation, whereas structurally similar but larger ice nucleation proteins (INPs) promote it (Fig. 1a, b)^{19–21}. Because graphene oxide (GO) nanosheets influence ice nucleation^{22–25} and can be prepared in a wide range of sizes, we used them to systematically explore the effect of nanoparticle size on ice nucleation.

GOs with different sizes were prepared by fractionating commercial GO aqueous dispersions by consecutively filtering through ultrafiltration membranes (Ultracel) with different molecular weight cut-offs (see Methods). Figure 1d–h shows transmission electron microscopy (TEM) images of GO fractions with average lateral sizes of 3 nm, 8 nm, 11 nm, 21 nm and 50 nm, respectively, along with the size distribution of each fraction. Atomic force microscopy (AFM) imaging indicates that the GOs have roughly the same thickness, irrespective of size (Extended Data Fig. 1a). Detailed characterizations of GOs with dynamic light scattering, X-ray photoelectron spectroscopy, Raman spectroscopy and nuclear magnetic resonance spectroscopy (see Extended Data Table 1

and Supplementary Figs. 1–5 for details) and cryo-TEM (Extended Data Fig. 1b) further consolidate that readily water-dispersible thin plate-like GOs with various sizes were obtained.

Ice nucleation activities were then probed by using optical microscopy to determine mean ice nucleation temperatures (T_{IN}) of water droplets containing GOs of different sizes (Methods). The top and bottom row images in Fig. 2a illustrate typical freezing behaviours seen in water droplets containing GOs with an average lateral size of 8 nm and 11 nm, respectively. Strikingly, the T_{IN} of the droplet containing 8-nm GOs is -27.6°C (Supplementary Video 1), which is about 10°C lower than that of a droplet containing 11-nm GOs under otherwise identical experimental conditions (Supplementary Video 2).

Figure 2b summarizes the results of our systematic exploration. We find that below 8 nm, T_{IN} is about -27.5°C and independent of GO size and concentration, and that it is equal to the T_{IN} measured under identical conditions for water droplets without added GOs. Because the homogeneous T_{IN} is lower than the T_{IN} that we see in this regime^{26,27}, we infer that ice formation is triggered by interfaces other than those of GO, for example the water–substrate interface (Extended Data Fig. 2b). When moving from GOs with a size of 8 nm to GOs with a size of 11 nm,

¹Key Laboratory for Green Printing, Beijing National Laboratory for Molecular Science, Institute of Chemistry, Chinese Academy of Sciences, Beijing, China. ²Research Institute for Energy Equipment Materials, School of Materials Science and Engineering, Hebei University of Technology, Tianjin, China. ³Key Laboratory of Hebei Province for Molecular Biophysics Institute of Biophysics, Hebei University of Technology, Tianjin, China. ⁴School of Physical Sciences and CAS Center for Excellence in Topological Quantum Computation, University of Chinese Academy of Sciences, Beijing, China. ⁵Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou, China. ⁶Songshan Lake Materials Laboratory, Dongguan, Guangdong, China. ⁷School of Future Technology, University of Chinese Academy of Sciences, Beijing, China. *e-mail: xzhou@ucas.ac.cn; wangjj220@iccas.ac.cn

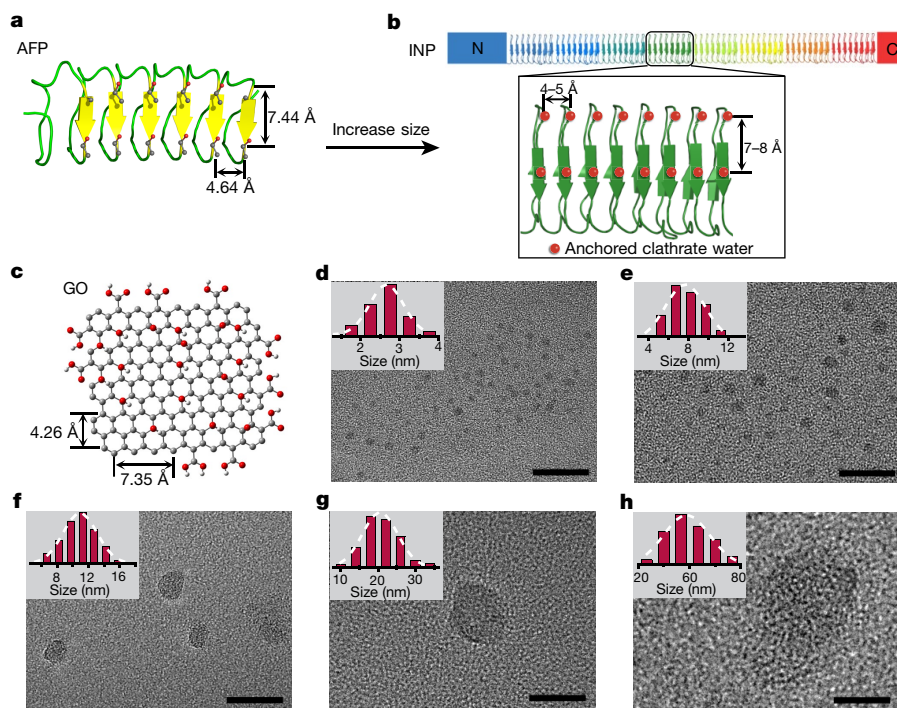


Fig. 1 | GOs of controlled sizes. **a**, Ribbon illustration of the antifreeze protein developed by the mealworm *Tenebrio molitor* (*TmAFP*). **b**, Schematic representation of monomer of ice nucleation protein in the bacterium *Pseudomonas syringae* (*PsINP*). The central tandem repeats of *PsINPs* have almost the same β -solenoid structure as that of *TmAFP*. Both proteins share a similar lattice feature with that of ice crystals^{19,20}. The main difference is that the central β -solenoid region of *PsINPs* is almost ten times as large as that of *TmAFP*. **c**, Illustration of GO nanosheets. Carbon, grey; oxygen, red; hydrogen, white. **d–h**, TEM images of various-sized GOs (see Methods). All the scale bars are 20 nm. The insets in **d–h** are the corresponding size distributions of the GOs. Each size distribution is obtained by analysing the lateral diameters of more than 100 GOs imaged by TEM.

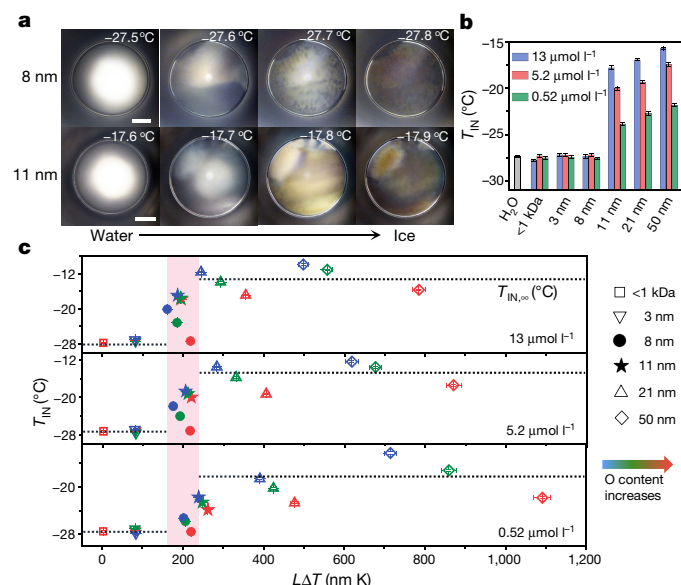


Fig. 2 | Ice nucleation activities of GOs with different sizes and oxidation degrees. **a**, Optical microscopic images showing typical freezing processes of water droplets (0.2 μl) containing GOs with average lateral sizes of 8 nm (upper row) and 11 nm (lower row) when the temperature was lowered at a cooling rate of 5 $^{\circ}\text{C min}^{-1}$. The GO concentrations in the water droplets are the same (13 $\mu\text{mol l}^{-1}$). Scale bar, 200 μm . **b**, T_{IN} of water droplets (0.2 μl) containing GOs of controlled sizes, at three different concentrations. Cooling rate, 5 $^{\circ}\text{C min}^{-1}$. Data are the mean \pm the standard error on the mean (s.e.m.). For each mean, the total number of measurements is about 150. **c**, The relationship between T_{IN} and $\Delta\Delta T$ (the supercooling scaled size of GOs, $\Delta T = T_{\text{m}} - T_{\text{IN}}$) for three different concentrations of GOs with six sizes and three oxidation extents. Data are means. The error bars for T_{IN} are s.e.m., and the error bars for $\Delta\Delta T$ are calculated according to the s.e.m. of L and T_{IN} based on the error propagation formulae. For each mean of T_{IN} or L , the total number of measurements is about 150.

we see an abrupt increase of about 10 $^{\circ}\text{C}$ in T_{IN} . The abrupt change persists when using different GO concentrations (Fig. 2b, and Extended Data Fig. 3), GOs with different degrees of oxidation (see Extended Data Table 2 and Supplementary Figs. 6–12 for details) and different cooling rates (Extended Data Fig. 4). Above 11 nm, further increases in GO size give rise to only slight further increases in T_{IN} .

The abrupt change in T_{IN} occurs at $\Delta\Delta T \approx 200$ nm K (Fig. 2c); here L is the average lateral size of GOs, and $\Delta T = T_{\text{m}} - T_{\text{IN}}$, with T_{m} being the equilibrium melting temperature of ice. When $\Delta\Delta T < 200$ nm K, ice nucleation occurs on the water–substrate interface and is little influenced by the presence of GOs. When $\Delta\Delta T > 200$ nm K, T_{IN} is almost independent of the value of $\Delta\Delta T$ but varies with GO concentration and corresponds to the normal heterogeneous ice nucleation temperature $T_{\text{IN},\infty}(C)$ associated with GOs large enough to induce ice nucleation. Note that we neglect the small changes in nucleation temperature associated with changes in the oxygen content of the investigated GOs.

The ice nucleation activity of GO sheets itself thus exhibits a transition when $\Delta\Delta T \approx 200$ nm K, which for GOs of any specific size L should occur at the supercooling temperature $\Delta T_L \approx (200 \text{ nm}/L)\text{K}$. As the degree of supercooling reached before heterogeneous ice nucleation sets in depends on the number, n , of contained GO sheets, we verify the expected change in GO ice nucleation activity by measuring T_{IN} for water droplets containing GOs from the same size fraction but in different numbers n (achieved by varying the concentration, the droplet volume or both). This largely excludes features unique to the differently sized GOs from influencing the ice nucleation trends that we see. As shown in Fig. 3a, we find for GOs with $L = 8$ nm, 11 nm and 21 nm respectively that T_{IN} increases with the logarithm of n only when $\Delta T > \Delta T_L$, revealing that at this temperature range, GO is active in facilitating ice nucleation; GO does not show an obvious effect on ice nucleation after $\Delta T \leq \Delta T_L$, as indicated by the fact that T_{IN} remains almost constant as the logarithm of n increases. The same behaviour is seen for GOs with three different oxygen contents (all with $L = 11$ nm) (Extended Data Fig. 5a).

We also measured ice nucleation delay times (t_{b}) as a function of supercooling (Fig. 3b), again finding a distinct change in t_{b} at the expected size-dependent supercooling $\Delta T_L \approx (200 \text{ nm}/L)\text{K}$. This was seen with the 8-nm, 11-nm and 21-nm GO samples, which were each used

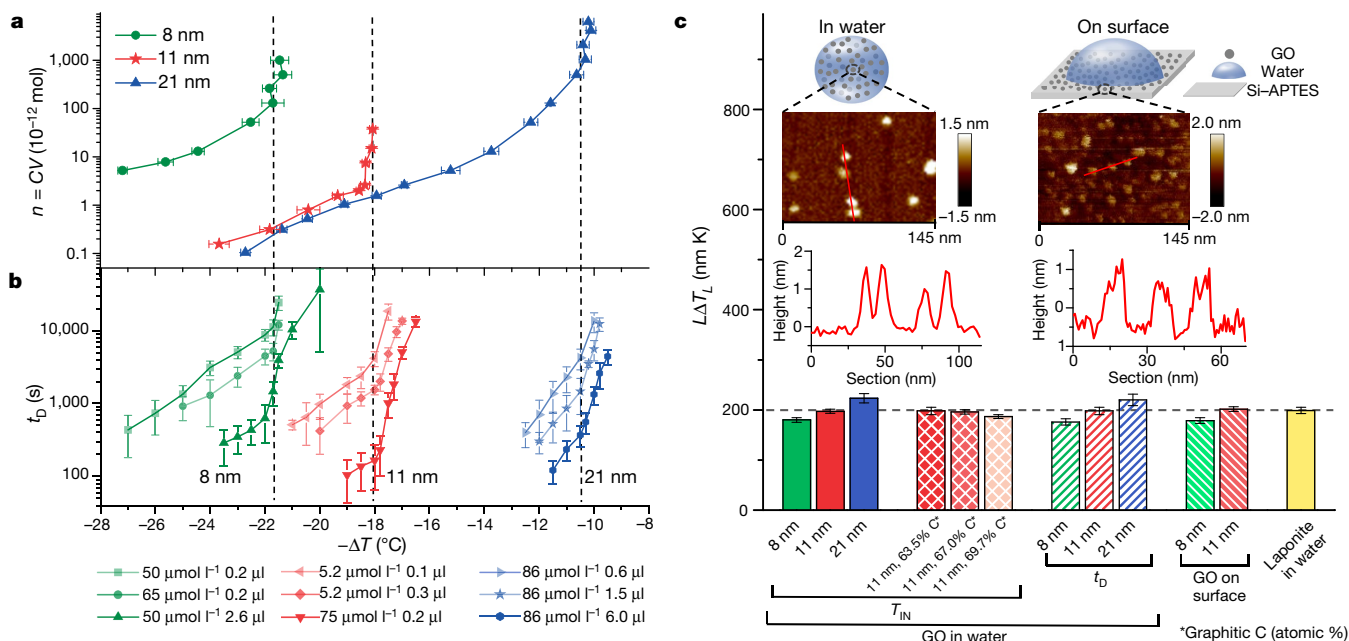


Fig. 3 | Transitions in the ice nucleation activity of nanosheets. a, The supercooling temperature of ice nucleation versus number of GOs in water droplet, $n = CV$, where C is the concentration of the GO aqueous dispersion and V is the volume of an individual droplet (data in Supplementary Table 1). Data are means \pm s.e.m. For each mean, the total number of measurements is about 50. **b**, The ice nucleation delay time for water droplets versus the supercooling ΔT . Data are means; error bars are standard deviation estimated by the jackknife resampling technique. For each mean, the total number of measurements varies from 20 to 150 to ensure that the nucleation event number m is typically not less than 10 (see Methods). **c**, The obtained ΔT_L for

anisotropic nanosheets. Here ΔT_L is the supercooling temperature at which the transition happens. Insets show schematic diagrams of water droplets containing GO nanosheets or water droplets deposited on the substrate anchored with GO nanosheets as well as the corresponding AFM images of GOs (see Methods), together with heights through the cross-section (obtained by AFM). Data are means; error bars of ΔT_L are calculated according to the s.e.m. of ΔT_L or L based on the error propagation formulae. For each mean of T_{IN} and L , the total number of measurements is about 50 and 100, respectively. Here the cooling rate is always 5°C min^{-1} .

with three different values of n , giving $\tau(T) = nt_D(T; n)$. The obtained $\tau(T)$ is independent of the number of GOs in aqueous dispersion within experimental error (Extended Data Fig. 5b), which agrees with the theoretical analysis (see Methods).

Figure 3c summarizes our experimental findings, illustrating that all investigated nanosheets exhibit an abrupt change in their ability to facilitate ice nucleation at ΔT_L , with a small deviation of only about 10%. This holds for ΔT_L inferred from different measurements (T_{IN} and t_b), different kinds of materials (GOs and laponite nanosheets, Extended Data Fig. 6) and different exposure of the nanosheets (either dispersed in water or anchored on a substrate, Extended Data Fig. 7). Note that measurements on the GO nanosheets anchored on solid surfaces exclude the possible influence on ice nucleation due to diffusion of GOs and the interplay among GO nanosheets when dispersed in water.

We can infer the free-energy barrier of ice nucleation (ΔG^*) from both $n(T)$ and $\tau(T)$ (see Methods), with Fig. 4a showing that the values collapse into the same line, that is, $\Delta G^* \propto \Delta T^{-2}$, over a small temperature range when $\Delta T > \Delta T_L$, consistent with classical nucleation theory (CNT). Importantly, ΔG^* shows an abrupt change at ΔT_L in all cases, revealing that the source of the abrupt change in ice nucleation activity of GOs is the change in the free-energy barrier for ice nucleation. The dependence of the free-energy barrier on the size of nanosheets is known to be based on the dimensionless variable, $l = L/(2R_c)$, that is, the relative size of the nanosheets to the radius R_c of the critical ice nucleus. Therefore, the transition of ΔG^* found experimentally to occur at the specific value of the dimensionless size of GO, $L_c/(2R_c) = l_c$, corresponds to $L_c \Delta T \approx 200 \text{ nm K}$; as such, we have $R_c = (100 \text{ nm K})/(l_c \Delta T) \propto \Delta T^{-1}$, consistent with CNT.

We explore this further by using CNT to calculate the free-energy barrier of ice nucleation on finite-sized GO nanosheets, consolidating that ΔG^* is a function of the dimensionless size of nanosheets and

has a transition at $l_c \approx 1$ almost regardless of the detailed features of nanosheets such as the shape and the interaction with ice (Fig. 4b; see Methods for more details). As sketched in Fig. 4c (and Extended Data Fig. 8), when $l < 2R_c$, two critical ice nuclei need to form in succession, and two corresponding free-energy barriers must be overcome. When $l > l_c$, the first free-energy barrier is the major one: the corresponding critical ice nucleus is a spherical cap with a small contact angle sitting on the surface of GO, the same as the heterogeneous ice nucleation atop GOs of sufficiently large size. By contrast, when $l < l_c$, the growing ice nucleus changes its shape after meeting the edge of GO and leads the second free-energy barrier to be the greater one. The corresponding critical ice nucleus is a spherical cap with a large contact angle due to the pinning at the edge of GO. Here the pinning is not due to any specific interaction of the edge of the GO with the water or ice, but is the requirement for minimizing the total interfacial free energy of the ice nucleus (see Methods); thus it is general. Therefore, the transition occurs when the major free-energy barrier alters from one to the other as l varies across the l_c at which the two free-energy barriers are equal.

Since $L_c/(2R_c) = l_c \approx 1$ from the theoretical calculation, we conclude that the critical size of GO, L_c , is approximately equal to the diameter of the critical ice nucleus, thus $R_c \approx (100/\Delta T) \text{ nm}$ (where ΔT is in kelvin). According to CNT, $R_c = 2\gamma/|\Delta\mu|$, and we can obtain the interfacial energy between ice and water to be $\gamma \approx 45 \text{ mJ m}^{-2}$, if using a typical value of the chemical potential difference between ice and water, $\Delta\mu \approx -893 \Delta T \text{ mJ cm}^{-3}$. Note that the value of γ cannot be directly measured experimentally, and the reported γ has a large range from 23 mJ m^{-2} to 54 mJ m^{-2} in the literature^{13,28,29}. The current method provides a way to measure the value of γ . Our results also show that surfaces with a pattern size comparable to that of the critical ice nucleus (for example,

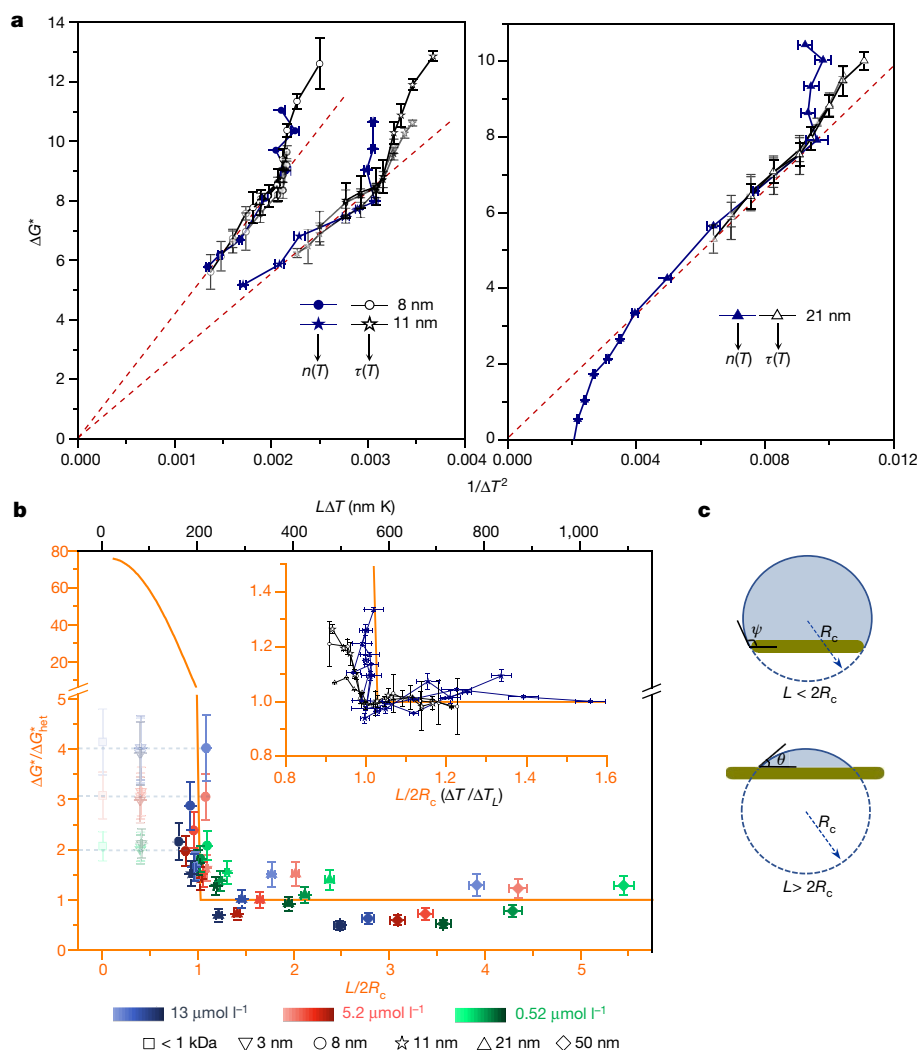


Fig. 4 | Abrupt change in the free-energy barrier of ice nucleation on GO nanosheets. **a**, The free-energy barrier ΔG^* (with units of $k_B T$) is obtained from the curves of $n(T)$ and $\tau(T)$, respectively (see Methods). The T_{IN} data for $n(T)$ here are medians \pm the standard error of the median (estimated as 1.2533 s.e.m.). For each median, the total number of measurements is about 50. The data for $\tau(T)$ are from Fig. 3b, including the data for all three different numbers of GOs (indicated by three different transparencies). The dashed line gives ΔG^*_{het} , the free-energy barrier of the normal heterogeneous nucleation on sufficiently large GOs. **b**, The free-energy barrier obtained from the data in Fig. 2c is compared with the one obtained from the CNT calculation (orange

line and axis) with typical assumptions of GO characteristics (see Methods). Data shown in paler colours ($\Delta T < 200$ nm K) correspond to nucleation on the water/substrate interface. Inset, the free-energy barriers in Fig. 4a are compared with those from the CNT calculation (orange). The error bars of the calculated parameters are calculated according to the error propagation formulae. **c**, Schematic illustrations of the shape of the critical ice nucleus when the size of GO nanosheet is smaller or larger than the critical ice nucleus diameter $2R_c$. ψ and θ are the apparent contact angles of the sphere-cap ice nucleus with the nanosheet plane.

surfaces anchored with nanosized GOs) can alter a surface's ability to control ice formation, which provides a strategy for the design of anti-icing surface materials³⁰.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1827-6>.

1. Sossio, G. C. et al. Crystal nucleation in liquids: open questions and future challenges in molecular dynamics simulations. *Chem. Rev.* **116**, 7078–7116 (2016).
2. Gallo, P. et al. Water: a tale of two liquids. *Chem. Rev.* **116**, 7463–7500 (2016).
3. Zhang, Z. & Liu, X. Y. Control of ice nucleation: freezing and antifreeze strategies. *Chem. Soc. Rev.* **47**, 7116–7139 (2018).
4. Kiselev, A. et al. Active sites in heterogeneous ice nucleation—the example of K-rich feldspars. *Science* **355**, 367–371 (2017).

5. He, Z., Liu, K. & Wang, J. Bioinspired materials for controlling ice nucleation, growth, and recrystallization. *Acc. Chem. Res.* **51**, 1082–1091 (2018).
6. Moore, E. B. & Molinero, V. Structural transformation in supercooled water controls the crystallization rate of ice. *Nature* **479**, 506–508 (2011).
7. Matsumoto, M., Saito, S. & Ohmine, I. Molecular dynamics simulation of the ice nucleation and growth process leading to water freezing. *Nature* **416**, 409–413 (2002).
8. Fitzner, M., Sossio, G. C., Pietrucci, F., Pipolo, S. & Michaelides, A. Pre-critical fluctuations and what they disclose about heterogeneous crystal nucleation. *Nat. Commun.* **8**, 2257 (2017).
9. Pereyra, R. G., Szleifer, I. & Carignano, M. A. Temperature dependence of ice critical nucleus size. *J. Chem. Phys.* **135**, 034508 (2011).
10. Pradzynski, C. C., Forck, R. M., Zeuch, T., Slavicek, P. & Buck, U. A fully size-resolved perspective on the crystallization of water clusters. *Science* **337**, 1529–1532 (2012).
11. Xiao, Q. et al. What experiments on pinned nanobubbles can tell about the critical nucleus for bubble nucleation. *Eur. Phys. J. E* **40**, 114 (2017).
12. Lupi, L., Peters, B. & Molinero, V. Pre-ordering of interfacial water in the pathway of heterogeneous ice nucleation does not lead to a two-step crystallization mechanism. *J. Chem. Phys.* **145**, 211910 (2016).
13. Cabriolu, R. & Li, T. Ice nucleation on carbon surface supports the classical theory for heterogeneous nucleation. *Phys. Rev. E* **91**, 052402 (2015).
14. Lupi, L. et al. Role of stacking disorder in ice nucleation. *Nature* **551**, 218–222 (2017).
15. Russo, J., Romano, F. & Tanaka, H. New metastable form of ice and its role in the homogeneous crystallization of water. *Nat. Mater.* **13**, 733–739 (2014).
16. Palmer, J. C. et al. Metastable liquid–liquid transition in a molecular model of water. *Nature* **510**, 385–388 (2014).

17. Fletcher, N. H. Size effect in heterogeneous nucleation. *J. Chem. Phys.* **29**, 572–576 (1958).
18. Welts, A., Lüönd, F., Stetzer, O. & Lohmann, U. Influence of particle size on the ice nucleating ability of mineral dusts. *Atmos. Chem. Phys.* **9**, 6705–6715 (2009).
19. Liou, Y. C., Tocilj, A., Davies, P. L. & Jia, Z. C. Mimicry of ice structure by surface hydroxyls and water of a beta-helix antifreeze protein. *Nature* **406**, 322–324 (2000).
20. Garnham, C. P., Campbell, R. L., Walker, V. K. & Davies, P. L. Novel dimeric beta-helical model of an ice nucleation protein with bridged active sites. *BMC Struct. Biol.* **11**, (2011).
21. Liu, K. et al. Janus effect of antifreeze proteins on ice nucleation. *Proc. Natl Acad. Sci. USA* **113**, 14739–14744 (2016).
22. Whale, T. F., Rosillo-Lopez, M., Murray, B. J. & Salzmann, C. G. Ice nucleation properties of oxidized carbon nanomaterials. *J. Phys. Chem. Lett.* **6**, 3012–3016 (2015).
23. Häusler, T. et al. Ice nucleation activity of graphene and graphene oxides. *J. Phys. Chem. C* **122**, 8182–8190 (2018).
24. Lupi, L., Hudait, A. & Molinero, V. Heterogeneous nucleation of ice on carbon surfaces. *J. Am. Chem. Soc.* **136**, 3156–3164 (2014).
25. Zheng, Y., Su, C., Lu, J. & Loh, K. P. Room-temperature ice growth on graphite seeded by nano-graphene oxide. *Angew. Chem.* **52**, 8708–8712 (2013).
26. Roscoe, R. B. How does a rain drop grow? *Science* **129**, 123–129 (1959).
27. Koop, T., Luo, B. P., Tsias, A. & Peter, T. Water activity as the determinant for homogeneous ice nucleation in aqueous solutions. *Nature* **406**, 611–614 (2000).
28. Li, T. S., Donadio, D., Russo, G. & Galli, G. Homogeneous ice nucleation from supercooled water. *Phys. Chem. Chem. Phys.* **13**, 19807–19813 (2011).
29. Némec, T. Estimation of ice–water interfacial energy based on pressure-dependent formulation of classical nucleation theory. *Chem. Phys. Lett.* **583**, 64–68 (2013).
30. Eberle, P., Tiwari, M. K., Maitra, T. & Poulidakos, D. Rational nanostructuring of surfaces for extraordinary icephobicity. *Nanoscale* **6**, 4874–4881 (2014).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Preparation of GOs of controlled sizes

The aqueous dispersion of GOs with a broad size distribution was procured from XFANO Materials Tech (Nanjing). It was size-fractionated using a stirred cell (Millipore Amicon) with an Ultracel membrane inside it under a pressure of about 0.4 MPa. GOs of various sizes can be obtained by using Ultracel membranes with molecular weight cut-offs of 1 kDa, 50 kDa, 100 kDa, 300 kDa and 500 kDa, and the 0.1- μm microfiltration membrane. Specifically, six GO fractions were obtained. The first fraction (<1 kDa) is the filtrate of the 1 kDa membrane (not shown). The other fractions were obtained by (for example) filtering commercial GOs through a membrane with a 50 kDa cut-off, and then removing small GOs in the filtrate by allowing them to pass through a membrane with a 1 kDa cut-off, retaining only the larger ones (Fig. 1d); and similarly obtaining filtered products of the 50 kDa and 100 kDa membranes (Fig. 1e); 100 kDa and 300 kDa membranes (Fig. 1f); 300 kDa and 500 kDa membranes (Fig. 1g); and 500 kDa and 0.1 μm membranes (Fig. 1h), respectively. The GO fraction obtained was kept in water and stored at 4 °C in a refrigerator when not in use.

The mass concentrations of these GO aqueous dispersions were measured by weighing the solid content of GO in a fixed volume of the dispersion. Specifically, a coverslip was first weighed by an analytical balance (with the accuracy of 0.01 mg); then a fixed volume (such as 500 μl) of dispersion was carefully dripped on the coverslip and was dried in an oven. The coverslip with the dried GO was further weighed after it was cooled to room temperature, and the dry GO mass was obtained from the mass difference. Every concentration was measured at least three times for the mean. Afterwards, GO dispersions with desired concentrations were prepared by diluting these mother dispersions with ultrapure water (18.2 M Ω cm) provided by Millipore Milli-Q apparatus and filtered through the 0.22- μm membrane.

Estimation of GO molar concentrations

Molar concentrations of GO aqueous dispersions were estimated from their mass concentrations and the molar mass of GOs. First, the molecular weight of GOs of different sizes was estimated based on a previous well-accepted structural model of GO^{31,32}; that is, GO with a size of 2.13 nm \times 2.46 nm has a chemical formula of $\text{C}_{240}\text{O}_{24}(\text{OH})_{24}(\text{COOH})_{12}$. Therefore, the molar mass of a GO nanosheet of unit size $M_{\text{GO},A}$ was calculated as $M_{\text{GO},A} = M[\text{GO model}]/A[\text{GO model}]$. Here $M[\text{GO model}]$ and $A[\text{GO model}]$ are the molar mass and area (in nm^2) of $\text{C}_{240}\text{O}_{24}(\text{OH})_{24}(\text{COOH})_{12}$, respectively. As revealed by the TEM imaging, the shape of the GOs is nearly circular. The average molar mass of GOs with a certain average lateral size $M_{\text{GO},L}$ was then calculated as $M_{\text{GO},L} = \pi(L/2)^2 \times M_{\text{GO},A}$. Here L is the average lateral dimension of GO measured from TEM images. Based on the average molar mass, the molar concentration of GO aqueous dispersion with a known mass concentration was estimated.

Preparation of GO samples with decreasing degree of oxidation

To obtain GOs with the same size but decreasing degrees of oxidation, the prepared GOs of controlled sizes were deoxidized by the facile alkali treatment method^{33,34}. Specifically, the GO aqueous dispersion (typically 0.2 mg ml^{-1} and 20 ml) with GOs within a specific size range was placed into vials in triplicate. The pH values of two of these dispersions were adjusted to 10 and 12 with 1 mol l^{-1} NaOH solution, respectively. The third dispersion was untreated. Then the three dispersions were stirred for 12 h. Subsequently, these three samples were purified to remove NaOH and other small molecules using Millipore Amicon stirred cell with an Ultracel membrane (molecular weight cut-off, 1 kDa) inside it. The purification processes were repeated three times to ensure that all impurities were removed. Finally, we added the required amount of ultrapure water according to the desired concentration into the Millipore Amicon stirred cell, to obtain GO aqueous dispersions with the

same size range but different degrees of oxidation. The GOs without alkali treatment and with the alkali treatments at pH = 10 and 12 were named R0, R1 and R2, respectively. See Supplementary Figs. 6, 7, 8 and 9 for the elemental content, hydrodynamic diameter, zeta potential and dispersibility characterizations.

Anchoring of GOs on silicon wafer surfaces

GOs were anchored on the Si wafer surface via the electrostatic adsorption between the amino groups of aminopropyltriethoxysilane (APTES) on the substrate and carboxyl groups of GOs. First, the Si wafer surface was modified with APTES³⁵. The Si wafer surface covered with APTES was soaked in the GO aqueous dispersion (2 $\mu\text{mol l}^{-1}$) for 12 h, and then ultrasonically cleaned in ethanol (5 s, 100 W, 40 kHz) and rinsed with ultrapure water, followed by flushing with nitrogen gas. The obtained sample was denoted Si-APTES-GO.

Preparation of laponite aqueous dispersion

Laponite RD (chemical formula $\text{Na}^{+}_{0.7}[(\text{Si}_8\text{Mg}_{5.5}\text{Li}_{0.3})\text{O}_{20}(\text{OH})_4]^{-0.7}$) with a purity over 99% was a gift from Huizhi Fine Chemical (Sihong). Laponite powder (total 2.25 g) was added stepwise into 225 ml ultrapure water at 65 °C under vigorous stirring. The amount each time added was about 0.2 g. Note that no additional laponite powder was added to the water until the dispersion turned clear. The entire addition period was about 1 h.

The preliminarily dispersed laponite was then filtered through a membrane filter with a pore diameter of 1 μm . The filtrate was then treated with ultrasonication for 3 h (40 kHz, 300 W, KQ-300DE ultrasonic cleaner, Kunshan Ultrasonic Instruments). After this, the laponite aqueous dispersion was poured into a stirred cell (Millipore Amicon) with an Ultracel membrane with 100-kDa cut-off inside it and filtrated under a pressure of about 0.4 MPa to remove the smaller nanosheets and other small molecules. Finally, ultrapure water was added to obtain 140 ml laponite aqueous dispersion. The mass concentration was measured by weighing the dry laponite mass in the dispersion of a fixed volume, and the molar concentration was then estimated based on the density (2.5 g cm^{-3}) and size (obtained by analysing AFM images, Extended Data Fig. 6) of laponite. The newly prepared laponite aqueous dispersion was used for ice nucleation measurements within 2 days to avoid the formation of possible aggregates.

Characterizations of GOs

The sizes of various GO samples were measured based on the images taken with the transmission electron microscopy (TEM, JEM-2100F, JEOL). AFM (Multimode 8, Bruker) was also used to investigate the morphology and thickness of GOs. The morphology of GOs in water was further examined by cryo-TEM (Tecnai Arctica, FEI). Specifically, the vitrified specimen was prepared in a closed chamber with 100% relative humidity and fixed temperature of 4 °C. First, a 3- μl droplet of GO aqueous dispersion (0.4 mg ml^{-1}) was dripped onto a perforated carbon film-supported grid held by tweezers and pre-equilibrated in the chamber. Excess dispersion was removed by blotting with a piece of filter paper for 4 s, producing a thin liquid film spanning the holes of the grid. The grid was then plunged into the liquid nitrogen to create the vitrified sample. Micrographs were recorded by a K2 Summit direct electron detector (Gatan) at a nominal magnification of 120,000 \times . Raman spectra were taken on a Raman spectrometer equipped with a 532-nm laser (LabRAM HR Evolution, HORIBA). The elemental content and chemical bonding were determined by X-ray photoelectron spectroscopy (XPS, ESCALab220i-XL, Thermo Fisher Scientific). Peak deconvolution with Gaussian curves of elements was accomplished by XPSPEAK 4.1 software. Zeta potentials of GO aqueous dispersions were measured by a Malvern Zetasizer (Nano ZS90, Malvern). Hydrodynamic diameters of various GO samples were measured by dynamic light scattering spectrometer (ALV/SP-125, ALV) equipped with a multi- τ digital time correlator (ALV-5000) and a He-Ne laser (22 mW, $\lambda = 632.8 \text{ nm}$).

The measurements were conducted at a scattering angle of 90°. All dispersions were filtrated through syringe filters with pore size of 0.45 µm before the measurements. All the measurements were performed at 25.00 ± 0.01 °C. The data obtained by dynamic light scattering reflect the size change of different GO fractions. Solid-state ¹³C high-power proton decoupling NMR spectra were acquired on a Bruker Avance III-400 spectrometer (100.38 MHz ¹³C, 399.16 MHz ¹H) after excitation with a 30° pulse and with a recycle delay of 15 s. A total of 15,360 scans were accumulated to obtain good signal-to-noise ratio. A 4-mm rotor and a spinning rate of 12 kHz were used. Peak deconvolution was accomplished by MestReNova software to separate the crowded peak.

Ice nucleation measurement

The ice nucleation temperature T_{in} and delay time were measured in a closed cell consisting of a rubber O-ring (height 2.0 mm, inner diameter 15 mm) sandwiched between two optical microscope cover glasses. Inside the closed cell, about 10 droplets of water or GO aqueous dispersions were placed atop a circular cover glass (Linkam 3930) using transferpettes. To minimize the influence of the substrate on the ice nucleation and to ensure that the freezing events of each water droplet are independent (Extended Data Fig. 2a), the circular cover glass was coated in advance with a silicone oil thin film about 40 µm thick (AR 1000 from Aldrich, which has a higher density than that of water, 1.09 g ml⁻¹ at 20 °C)³⁶. The entire preparation of the sample cell was carried out in a Class II Type A2 biosafety cabinet to avoid contamination. All the water used in the experiments was ultrapure water. The closed cell is small enough (0.35 cm³) that the water vapour in the closed cell can be approximated to be 100% relative humidity. Then the closed cell was placed atop a cryostage (Linkam LTS420) and cooled at a rate of 1, 5, 10 or 15 °C min⁻¹.

The formation of ice was observed through an optical microscope (Nikon AZ100) equipped with a digital camera (Nikon DS-Ri1). The temperature at which a sudden change in the opacity of water droplets was first observed was identified as T_{in} . One-way analysis of variance (ANOVA) was also performed on the T_{in} data of water droplets containing GOs with a series of sizes for statistical significance (significance level of difference of the mean is 0.05; see Supplementary Section PS2).

For GOs anchored on Si wafer surfaces, T_{in} was measured in a similar way. The difference was that we replaced the silicone oil coated cover glass with the sample to be tested and then pure water droplets were placed atop the Si wafer surfaces anchored with GOs. The number of nucleation sites was tuned by the contact area of the water droplets with the substrate, achieved by changing the volumes of the water droplets. Every sample with water droplets atop was photographed by an optical microscope equipped with a digital camera, and then the images were analysed by the NIS-Elements BR software to obtain the contact area of the water droplet with the substrate.

The delay time of ice nucleation at a certain temperature was measured as the time elapsed from the time when the substrate was cooled to a target temperature to the time when the ice nucleation occurred. Estimate of mean delay time of ice nucleation was as follows. We independently measured the ice nucleation delay time N times in our experiments, and the longest waiting time was $t_0 = 9,000$ s. Within 9,000 s, we found m nucleation events at t_1, t_2, \dots, t_m , respectively, and the remaining $N - m$ measurements did not have nucleation events (N varies from 20 to 150 to ensure that m is typically not less than 10). We have an estimator of the delay time (the mean waiting time), $t_D = (1/m)[\sum_{i=1}^m t_i + \sum_{j=m+1}^N t_0]$. We applied the jackknife resampling technique to obtain the error of the estimator of the delay time.

The free-energy barrier from the ice nucleation temperature and the mean delay time

Generally, the ice nucleation rate, $J(T) = nK(T)\exp[-\Delta G^*/(k_B T)]$, determines the temperature of ice nucleation in the cooling

experiments and the mean delay time of ice nucleation t_D at each specific temperature. Here n is the number of ice nucleation active sites (GOs) in water droplets; $K(T)$ is the kinetic prefactor; ΔG^* is the major (highest) free-energy barrier of ice nucleation (if multiple barriers exist); k_B is the Boltzmann constant.

The ice nucleation temperature. When water droplets are slowly cooled, the probability that an ice nucleation event happens for the first time at temperature T is $P(T) = (1/\alpha)J(T)\exp[(1/\alpha)\int^T J(T')dT']$. Here $\alpha = |dT/dt|$ is the cooling rate. When T decreases, the ice nucleation rate $J(T)$ quickly increases; but the exponential term quickly decreases; thus $P(T)$ is significantly non-zero only in a small supercooling temperature range, corresponding to the detected nucleation temperature. Within a small temperature range, we approximately have $J(T) \approx nK\exp(-\Delta G^*)$, where K is constant, and $\Delta G^* = \Delta G^*/(k_B T)$.

We can define the mean temperature $\bar{T} = \int P(T)dT$, and the temperature T_f^x at which the cumulative probability of ice nucleation is x , satisfies $\int_{T_f^x}^{\bar{T}} J(T)/\alpha dT = \ln(1-x)$. Usually we set $x = 0.5$, then $T_f^{0.5}$ is the median temperature of ice nucleation. The mean temperature and the median temperature $T_f^{0.5}$ do not equal each other, but their difference is found to be very small in the current experiments (Supplementary Fig. 13). Thus, we usually do not distinguish them if not explicitly mentioned.

The relationship between $T_f^{0.5}$ and n satisfies the equation $J(T_f^{0.5}) = \alpha(\ln 2) \frac{d \ln n}{dT} \cdot \frac{dT}{dT^{0.5}}$. Since $\frac{d}{dT} \ln n(T)$ usually varies much more slowly with T in comparison with $J(T)$, we approximate it as a constant; thus we have $J(T_f^{0.5})$ being a constant when α is fixed. We can then determine the dependence of $T_f^{0.5}$ on n .

Therefore, we have $\ln n(T) \approx \Delta \tilde{G}^*(T) + c'$. Here c' is almost a constant when $\Delta T > \Delta T_L$, but changes with T when $\Delta T \leq \Delta T_L$, since $\frac{d}{dT} \ln n(T)$ is a constant when $\Delta T > \Delta T_L$, but changes rapidly with T when $\Delta T \leq \Delta T_L$.

The delay time of ice nucleation. The distribution of the delay times of ice nucleation at a fixed temperature is $P(t) = J(T)\exp[-tJ(T)]$. Thus, the mean delay time is $t_D = J(T)^{-1} = n^{-1}K^{-1}\exp(\Delta \tilde{G}^*)$ where K is the prefactor of the nucleation rate. Then we have $\ln \tau(T) \approx \Delta \tilde{G}^*(T) + c$ with constant c . Here $\tau(T) = nt_D(T; n)$ is independent of the number of GOs, n , since t_D is inversely proportional to n .

When $\Delta T > \Delta T_L$ and within a small range of temperature, we found that both $\ln n(T)$ and $\ln \tau(T)$ are linearly related to $1/\Delta T^2$ with different additional constants, consistent with CNT. Thus we fitted the free-energy barrier by $\Delta \tilde{G}^* \propto 1/\Delta T^2$ from the curves $\ln n(T)$ and $\ln \tau(T)$, as shown in Fig. 4a.

Ice nucleation on gold nanoparticles

We also investigate the ice nucleation of water droplets containing gold nanoparticles of controlled size and show the results in Supplementary Section PS5. Abrupt transition in the activity of the nanoparticles in facilitating ice nucleation occurs at a critical size of gold nanoparticle.

Theoretical calculation of free-energy barrier of ice nucleation on finite-sized GOs

Based on CNT, on sufficiently large GO surfaces, the free-energy barrier $\Delta \tilde{G}^* = \Delta \tilde{G}_{het}^* = \tilde{a}/\Delta T^2$, where $\tilde{a} = \frac{16\pi\gamma^3}{3|\Delta S|^2 k_B T} f(\theta)$ is approximately constant

if the temperature is limited within a small range. Here γ is the surface tension of the ice–water interface, ΔS is the entropy difference between ice and water at the equilibrium melting temperature, and $f(\theta)$ describes the capability of sufficiently large GOs in facilitating ice nucleation.

When the size of GOs is comparable with that of the critical ice nucleus, the free-energy barrier $\Delta \tilde{G}^*(L; \Delta T) = \Delta \tilde{G}_{het}^*(\Delta T) \hat{g}(l)$. Here $\hat{g}(l)$ is a function of the dimensionless size of GOs,

$l \equiv L/(2R_c)$ ($=L\Delta T/x_c \equiv \Delta T/\Delta T_l$), and $R_c = 2\gamma/(|\Delta S|\Delta T)$ is the radius of the critical ice nucleus. $x_c = 4\gamma/|\Delta S|$ is approximately constant, and $\Delta T_l = x_c/L$.

The function $\hat{g}(l)$ can be calculated by modelling the shape of GOs (see Supplementary Section PS6). Here we suppose that the GO is a thin flat disk with a smooth semi-circular edge. As shown in Extended Data Fig. 8a, $\hat{g}(l)$ has an abrupt transition at $l = l_c \approx 1$. The result is not sensitive to the detailed shape of nanosheet as discussed below. The free-energy barrier is determined by the shape of critical ice nucleus under the requirement of minimizing its total interfacial free energy, involving that of its GO-covered surface and that of the GO-uncovered surface (that is, the interface between water and ice). For large GOs, it results in a critical ice nucleus in the shape of a sphere-cap atop the flat surface of GO, with radius R_c regardless of GO, and the contact angle θ determined by the Young's equation, $\gamma \cos \theta = \gamma_{\text{WG}} - \gamma_{\text{IG}}$. Here γ_{WG} and γ_{IG} are the surface energy of the water–GO and ice–GO interfaces, respectively. For small GOs, it forces a complete covering of the ice nucleus on the flat surface of GO to minimize the interfacial free energy of the GO-covered surface of ice nucleus, $(\gamma_{\text{IG}} - \gamma_{\text{WG}})S_{\text{IG}} = (-\gamma \cos \theta)S_{\text{IG}}$ with $S_{\text{IG}} \approx (\pi/4)L^2$, and a partial spherical surface of the GO-uncovered water/ice surface of the ice nucleus. Thus, the critical ice nucleus is a sphere-cap pinned at the edge of GO, with approximately the same radius R_c , and a large contact angle ψ , with $\cos \psi \approx -\sqrt{1-l^2}$, almost regardless of the details of the edge of the GO nanosheets. Therefore, the free-energy barrier is $\Delta G^* = \gamma S_{\text{WI}} - (\gamma \cos \theta)S_{\text{IG}} - |\Delta \mu|V$. Here S_{WI} and V are the area of the water–ice interface and the volume of the critical ice nucleus, respectively; $S_{\text{WI}} \approx 2\pi R_c^2(1 - \cos \psi)$ and $V \approx (4\pi/3)R_c^3 f(\psi)$. Then $\hat{g}(l) \approx [1/f(\theta)][f(\psi) - (3/4)l^2[\cos \psi - \cos \theta]]$ has a similar abrupt change when $l = l_c \approx 1$.

We illustrate the change in shape of the ice nucleus during its growth on GO nanosheet. As shown in Extended Data Fig. 8b, when $L \approx 2R_c$, the first critical ice nucleus with radius R_c forms on the nanosheet surface because of thermodynamic fluctuation; and then the ice nucleus spontaneously grows until it meets the edge of the nanosheet. After that, the growing ice nucleus increases its contact angle and first decreases, then increases its radius because it is pinned at the edge of the nanosheet. This leads to the second critical ice nucleus which has almost the same radius R_c , but a larger contact angle. Extended Data Fig. 8c shows the changes of the free energy, contact angle and radius of ice nucleus with volume.

Comparison between the experimental and theoretical free-energy barriers

From the data shown in Fig. 2c, we calculate $\hat{g}(l) = \Delta T^2(L\Delta T; C)/\Delta T_\infty^2(C)$, where $l = L\Delta T/x_c$ with $x_c = 200$ nm K, and $\Delta T_\infty(C) = T_m - T_{\text{IN},\infty}(C)$ for each concentration of GO, C . From the data in Fig. 4a, we get $\hat{g}(l) = \Delta G^*(\Delta T)/\Delta G_{\text{het}}^*(\Delta T)$ (where the subscript 'het' means heterogeneous nucleation), with $l = \Delta T/\Delta T_l$ for GOs of various sizes by using the experimental value of their ΔT_l , respectively, and the free-energy barrier of the normal heterogeneous nucleation on an infinite plane substrate, $\Delta G_{\text{het}}^*(\Delta T) \propto 1/\Delta T^2$, when ΔT is near ΔT_l . The comparison is shown in Fig. 4b of the main text.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

31. Tu, Y. et al. Destructive extraction of phospholipids from *Escherichia coli* membranes by graphene nanosheets. *Nat. Nanotechnol.* **8**, 594–601 (2013).
32. Geng, H. et al. Graphene oxide restricts growth and recrystallization of ice crystals. *Angew. Chem.* **56**, 997–1001 (2017).
33. Rourke, J. P. et al. The real graphene oxide revealed: stripping the oxidative debris from the graphene-like sheets. *Angew. Chem.* **50**, 3173–3177 (2011).
34. Fan, X. et al. Deoxygenation of exfoliated graphite oxide under alkaline conditions: a green route to graphene preparation. *Adv. Mater.* **20**, 4490–4493 (2008).
35. Bai, G. et al. Self-assembly of ceria/graphene oxide composite films with ultra-long antiwear lifetime under a high applied load. *Carbon* **84**, 197–206 (2015).
36. Du, N., Liu, X. Y. & Hew, C. L. Ice nucleation inhibition—mechanism of antifreeze by antifreeze protein. *J. Biol. Chem.* **278**, 36000–36004 (2003).

Acknowledgements The work is supported by the National Natural Science Foundation of China through grant nos 21733010, 11574310 and 21534007, the National Key R&D Program of China 2018YFA0208502, and the Strategic Priority Research Program of Chinese Academy of Sciences, grant no. XDB28000000. We thank B. Guan and Y. Liu for help in cryo-TEM experiments.

Author contributions G.B., X.Z. and J.W. conceived the project and designed the experiments. G.B., D.G. and Z.L. performed the experiments. G.B., D.G., Z.L., X.Z. and J.W. analysed the data. G.B., D.G., X.Z. and J.W. prepared the manuscript.

Competing interests The authors declare no competing interests.

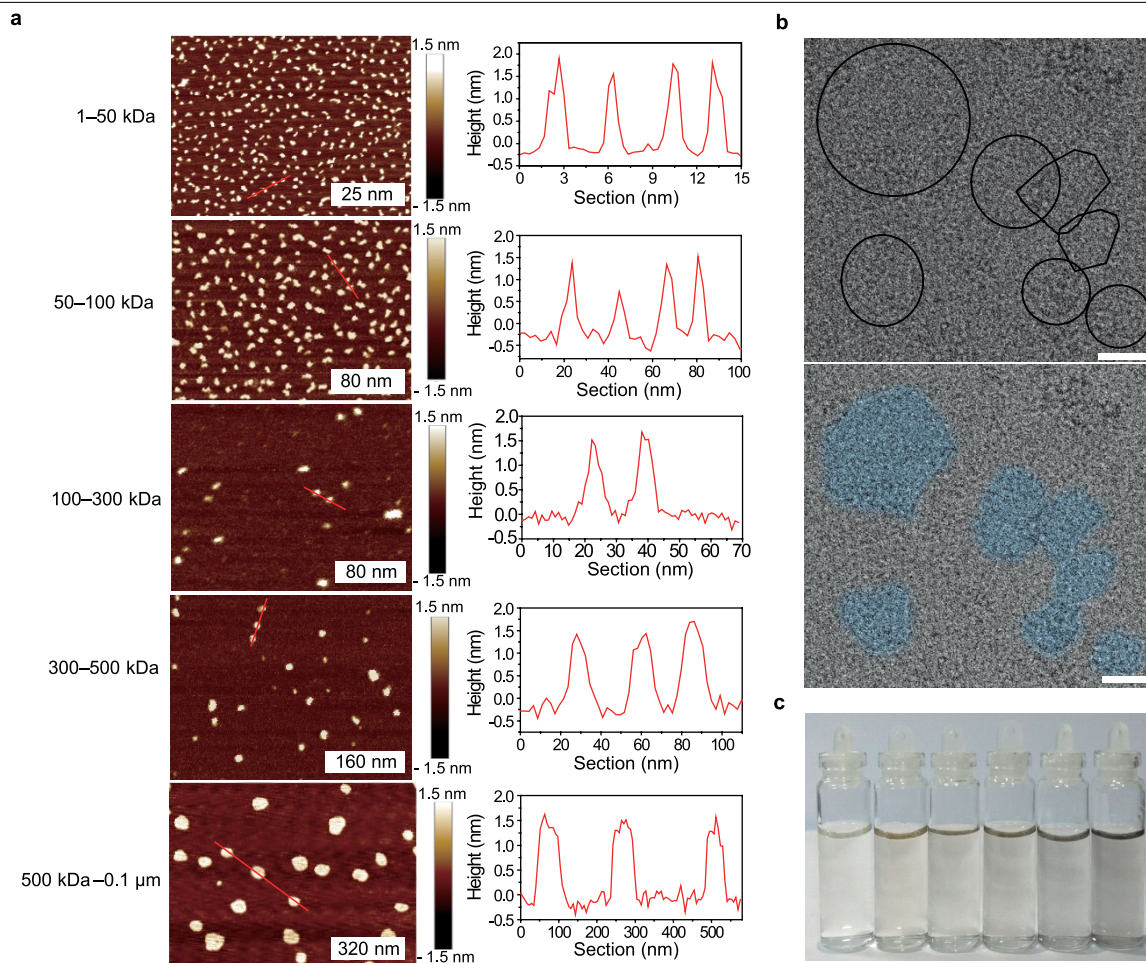
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1827-6>.

Correspondence and requests for materials should be addressed to X.Z. or J.W.

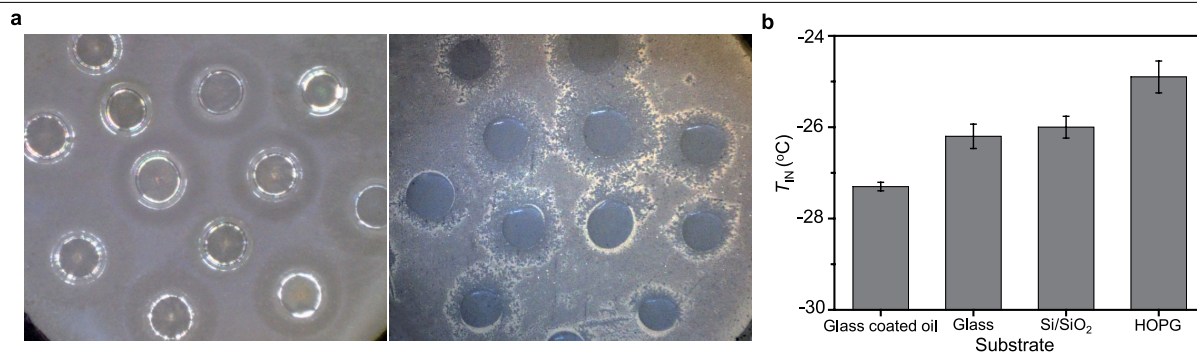
Peer review information Nature thanks Niall English, Christoph Salzmann and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



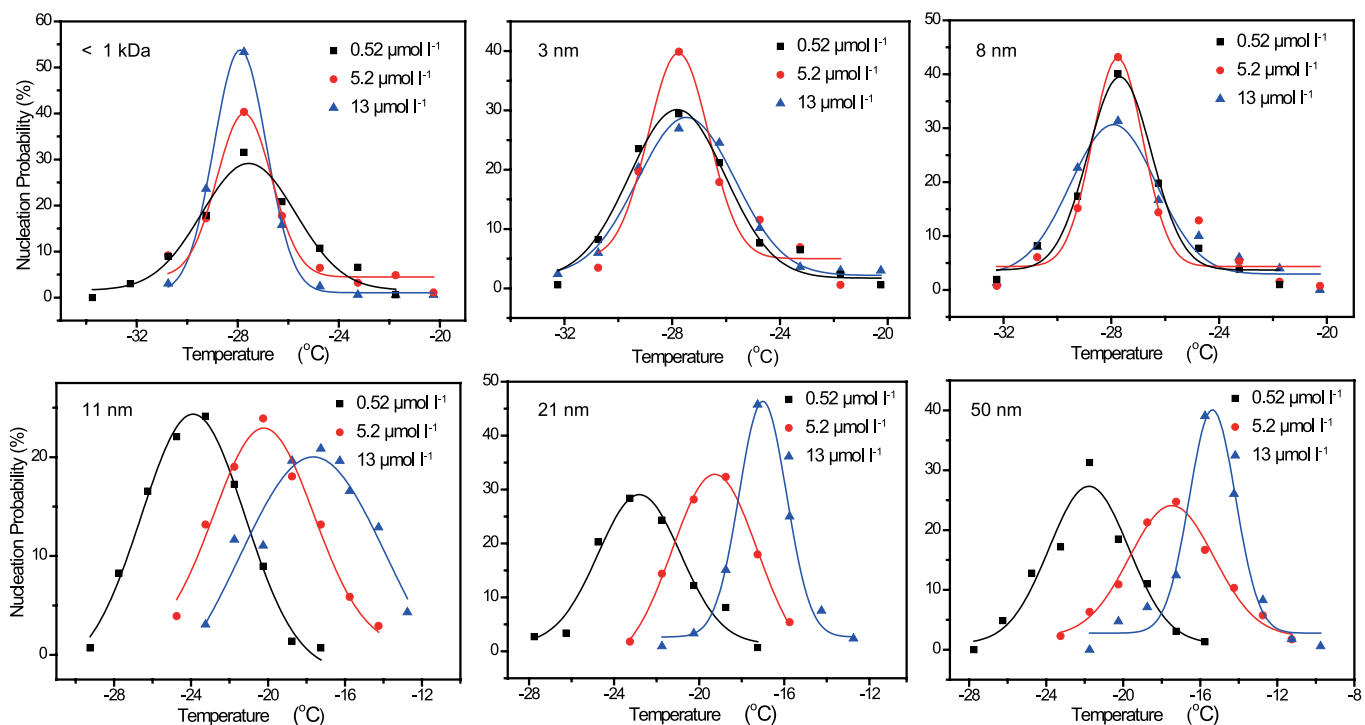
Extended Data Fig. 1 | Characterizations of GOs of controlled sizes. a, AFM images of GOs of five controlled sizes and the corresponding height profiles along the lines marked. **b**, Cryo-TEM images of GOs of various sizes before size fractionation, showing the shape of the GOs in water. The upper image is the original; the lower panel is the image with enhanced contrast by colouring the

GO domains to help the visibility. Scale bar, 10 nm. **c**, Photographs of 0.04 mg ml^{-1} GO aqueous dispersions. From left to right, the average lateral sizes of GO are <1 kDa, 3 nm, 8 nm, 11 nm, 21 nm and 50 nm, respectively. All the GO aqueous dispersions are clear and transparent, indicating the good dispersibility of various-sized GOs in water.

**Extended Data Fig. 2 | Influence of the substrate on the ice nucleation**

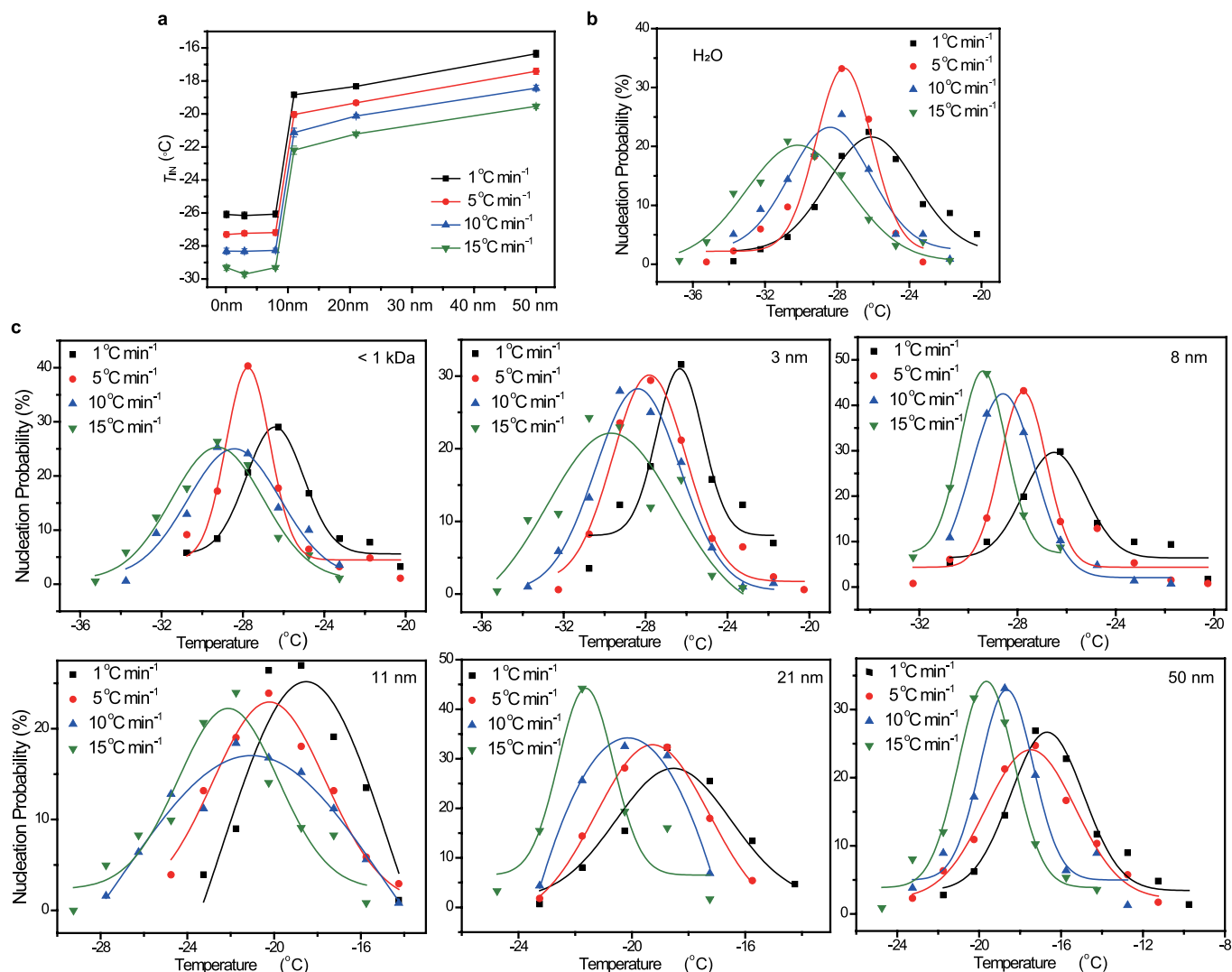
measurement. a, Optical microscopic images of frozen water droplets on glass coverslip coated with a thin layer of silicone oil (left) and without silicone oil (right) during the ice nucleation assays. The other experimental conditions for these two images are identical (see Methods). The frozen water droplets on a glass coverslip coated with a thin oil film are independent. In contrast, on the glass coverslip without a thin oil film, the freezing events of the water droplets

are not independent. **b,** Ice nucleation temperatures of water droplets on glass coated with silicone oil, glass without oil, silicon wafer and highly oriented pyrolytic graphite (HOPG). Data are means \pm s.e.m. For each mean, the total number of the measurements is not less than 50. The volume of the water droplet is 0.2 μ l. Cooling rate, 5 $^{\circ}\text{C min}^{-1}$. T_{IN} of water droplets on different substrates shows different values, suggesting that the ice nucleation is initiated at the water/substrate interface.



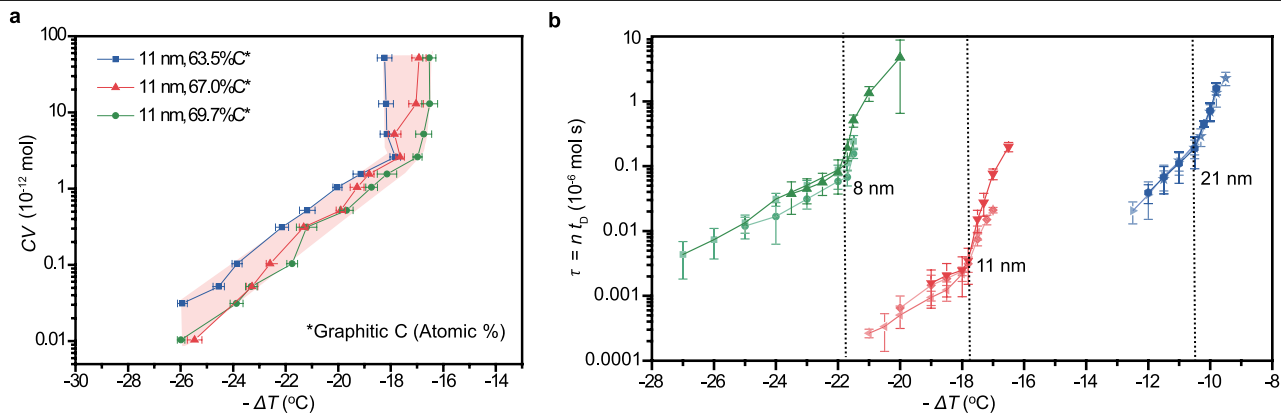
Extended Data Fig. 3 | Ice nucleation probability distribution of water droplets containing GOs of controlled sizes at three different concentrations. The distributions are fitted by Gaussian functions. For each distribution, the total number of ice nucleation measurements is about 150.

The results show that the change in concentration (from 0.52 to 13 $\mu\text{mol l}^{-1}$) of GOs with sizes smaller than 8 nm does not affect the T_{IN} of water droplets; however, T_{IN} increases with the concentration of GOs when the GO size is above 11 nm.



Extended Data Fig. 4 | Ice nucleation temperatures of droplets of GO aqueous dispersions at cooling rates ranging from 1 °C min⁻¹ to 15 °C min⁻¹. **a**, Cooling rate dependence of T_{IN} of water droplets containing GO samples of controlled sizes. Data are means \pm s.e.m. For each mean, the total number of measurements is about 150. **b**, Ice nucleation probability distribution of the blank control (water droplets) at various cooling rates with Gaussian fitting.

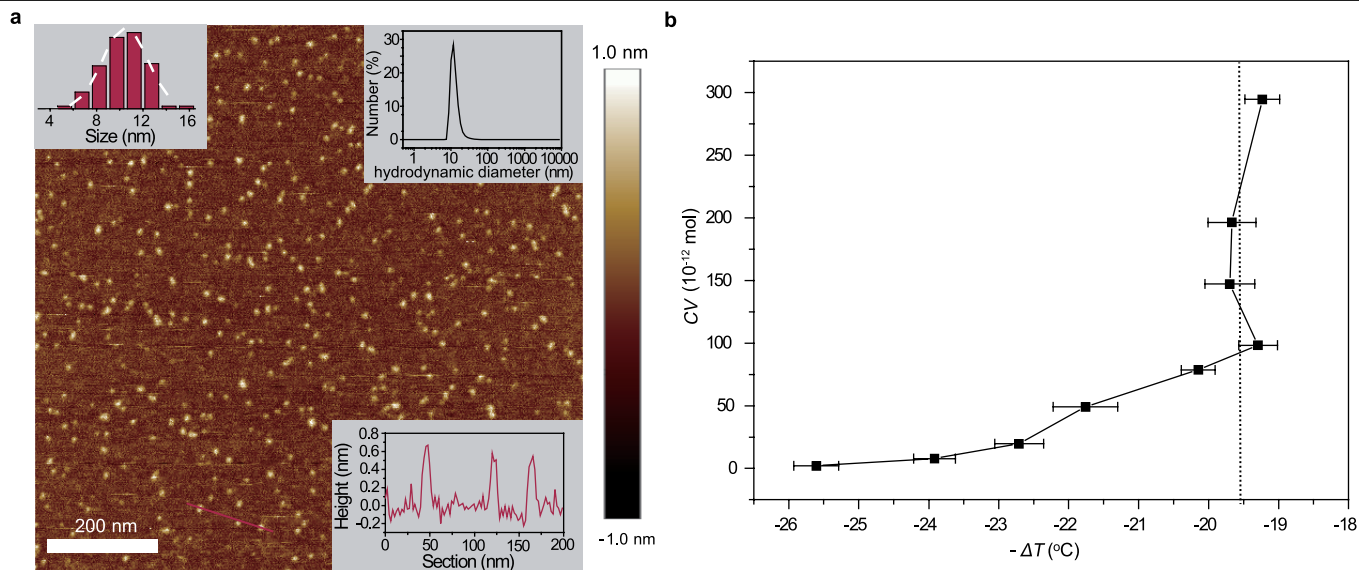
c, Ice nucleation probability distribution (Gaussian fitting) of water droplets containing GOs with a series of average lateral sizes at various cooling rates. For each distribution, the total number of measurements is about 150. The concentration of GO aqueous dispersion is 5.2 μ mol l⁻¹. All the volumes of water droplets are 0.2 μ l.



Extended Data Fig. 5 | The transitions of the ice nucleation activity of GOs.

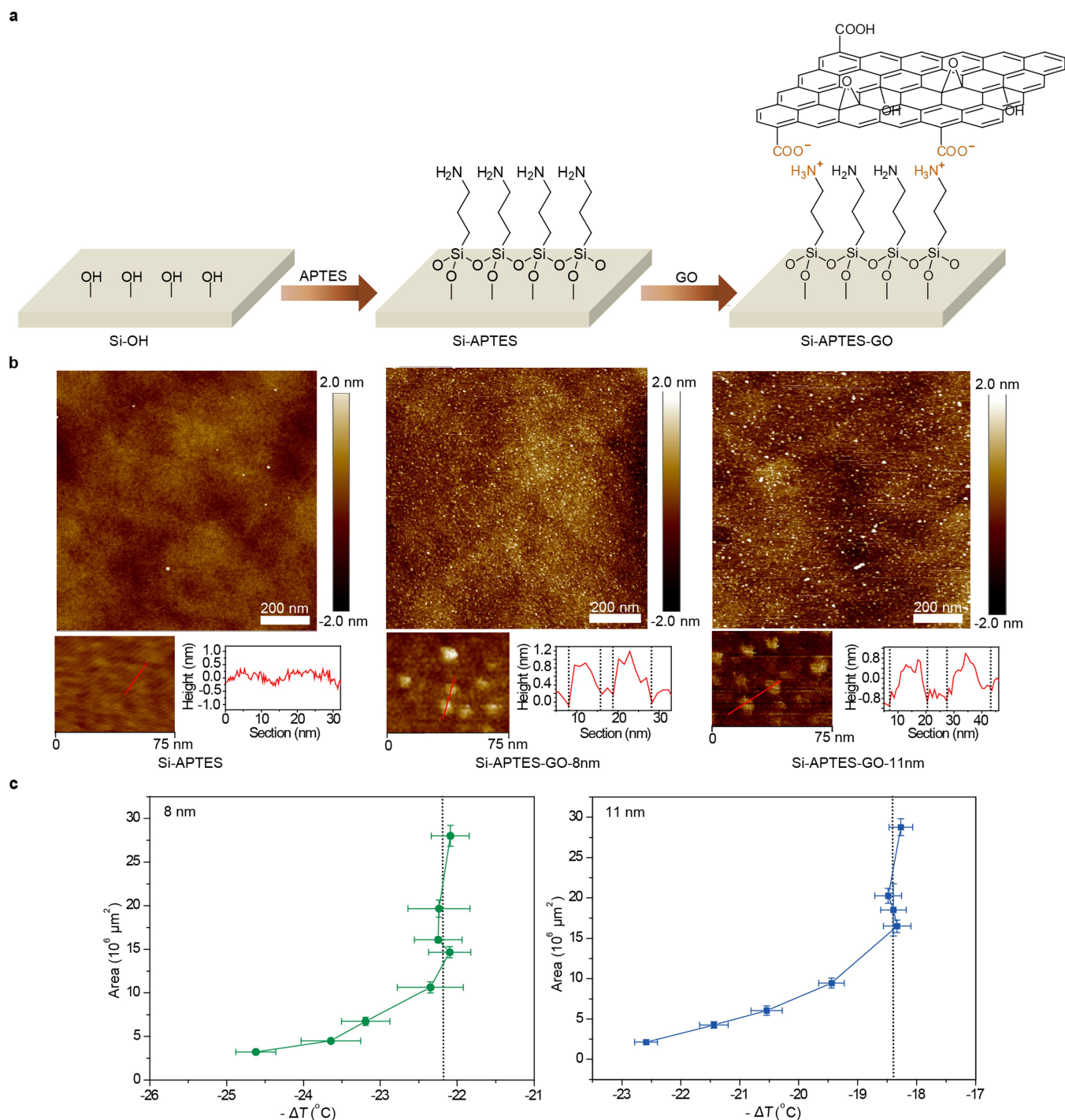
a, The mean ice nucleation (supercooling) temperature $-\Delta T \equiv T_{\text{IN}} - T_{\text{m}}$ versus the number of GOs in the water droplet, $n = CV$ (for concentration, C and volume, V), for three degrees of oxidation of GOs with the same lateral size of 11 nm. Here the cooling rate is always $5^{\circ}\text{C min}^{-1}$. Data are means \pm s.e.m. For each mean, the total number of the measurements is about 50. **b**, The scaled delay

time of ice nucleation of water droplets containing GOs, $\tau = n t_0(T; n)$, versus ΔT . The three curves for each GO size come from different n (the same as Fig. 3b in the main text) and collapse into the same curve. Data are means; error bars are standard deviation estimated by the jackknife resampling technique. For each mean, the total number of measurements varies from 20 to 150 to ensure that the nucleation event number m is typically not less than 10 (see Methods).



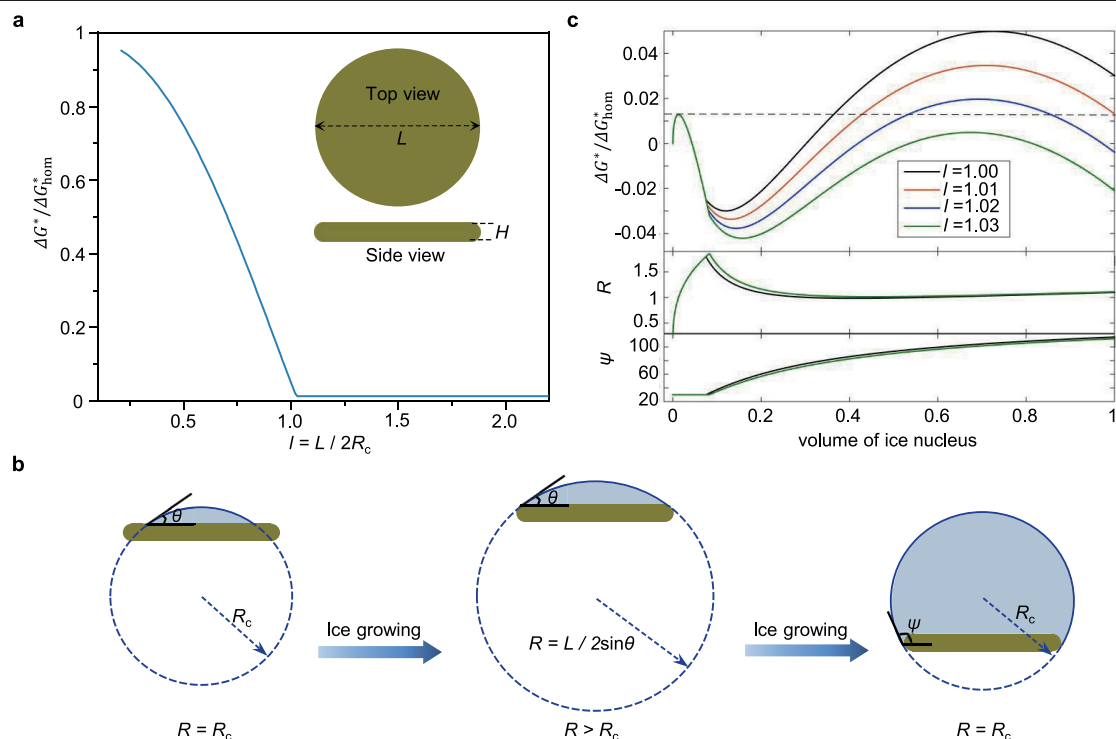
Extended Data Fig. 6 | Characterization and ice nucleation activity of laponite. a, AFM characterization of the prepared laponite. The insets show the lateral size distribution, the thickness and the hydrodynamic diameter of laponite. The size distribution is obtained by averaging the lateral sizes of more than 100 laponite nanosheets imaged by AFM. The hydrodynamic diameter of

laponite nanosheets is measured by a Malvern Zetasizer. **b,** The ice nucleation (supercooling) temperature $-\Delta T \equiv T_{IN} - T_m$ versus the number of laponites contained in the water droplet (for concentration, C and volume, V). Here the cooling rate is always 5°C min^{-1} . Data are means \pm s.e.m. For each mean, the total number of the measurements is about 50.



Extended Data Fig. 7 | Characterization and ice nucleation temperature investigations of GOs anchored on silicon wafer surface. **a**, Schematic illustration showing the preparation process of the anchored GOs on Si wafer surfaces. **b**, AFM characterizations of the prepared surfaces without GOs and with GOs of controlled sizes. **c**, The ice nucleation (supercooling) temperature

$-\Delta T \equiv T_{\text{IN}} - T_{\text{m}}$ versus the contact area between the water droplets and the surface to which the GOs are anchored. The contact area, measured by optical microscopy, is proportional to the number of nucleation active sites (see Methods). Here the cooling rate is always $5^{\circ}\text{C min}^{-1}$. Data are means \pm s.e.m. For each mean, the total number of measurements is about 50.



Extended Data Fig. 8 | Theoretical analysis of ice nucleation on finite-sized nanosheet. **a**, Free-energy barrier of ice nucleation on a thin-disk GO versus the normalized size of GOs. The inset shows the schematic illustration of thin-disk-shaped GOs with a smooth hemispherical edge. Its major diameter (lateral size) is L , and the thickness is H . **b**, Schematic diagram showing three typical shapes of ice nucleus on GO when $L \approx 2R_c$. The first and the third are the critical ice nuclei corresponding to two different free-energy barriers (see Methods).

c, The calculated dimensionless free energy, radius of ice nucleus (in units of R_c) and the apparent contact angle ψ versus the volume of ice nucleus (in units of $(4\pi/3)R_c^3$) on the thin-disk GO nanosheet when $L \approx 2R_c$. Here the dimensionless thickness of GO disk $h = H/2R_c = 0.1$, and $\theta (=30^\circ)$ is the intrinsic contact angle between ice nucleus and the GO. The obtained results are not sensitive to these details of GO and the applied parameters (see Methods and Supplementary Section PS6).

Extended Data Table 1 | Summary of characterization of GOs of controlled sizes

Sample	< 1 kDa	1–50 kDa	50–100 kDa	100–300 kDa	300–500 kDa	500 kDa–0.1 μm
TEM diameter (mean \pm s.d.) (nm)	N/A	2.63 \pm 0.46	7.63 \pm 1.54	10.91 \pm 2.2	20.91 \pm 4.41	49.75 \pm 12.11
Hydrodynamic diameter (mean \pm s.d.) (nm)	0.62 \pm 0.05	2.80 \pm 0.58	7.73 \pm 1.15	14.36 \pm 0.84	29.69 \pm 6.49	68.34 \pm 10.12
Zeta potential (mean \pm s.d.) (mV)	N/A	- 24.0 \pm 1.0	- 24.6 \pm 2.0	- 24.0 \pm 0.5	- 23.3 \pm 2.7	- 21.5 \pm 2.2
C/O atomic ratio	1.35	2.45	2.58	2.60	3.13	3.14
Graphitic C/oxidized C	1.48	1.54	1.57	1.67	1.85	2.01
Graphitic C (Atomic %)	59.6	60.7	61.1	62.8	64.9	66.8
C-O (Atomic %)	21.8	20.6	21.8	23.0	26.5	24.5
C=O/COOH (Atomic %)	18.6	18.7	17.1	14.2	8.6	8.7
I_D/I_G	1.19 \pm 0.07	1.12 \pm 0.01	1.05 \pm 0.02	1.02 \pm 0.01	1.01 \pm 0.02	1.01 \pm 0.03

The TEM diameter of each sample is obtained by averaging the lateral sizes of more than 100 GOs imaged by TEM. Hydrodynamic diameter and zeta potential for each sample are obtained by averaging three measurements. Hydrodynamic diameter and zeta potential distributions of GOs are shown in Supplementary Figs. 1 and 2. The carbon (C) content in different chemical states can be obtained from the area ratio of the sub-peaks in the C 1s core-level XPS spectra (Supplementary Fig. 3). I_D/I_G represents the intensity ratio of the D band to G band obtained from Raman spectroscopy (Supplementary Fig. 5) and is obtained by averaging three measurements.

Extended Data Table 2 | Ice nucleation temperatures of water droplets containing GOs of controlled sizes and decreasing degrees of oxidation

C ($\mu\text{mol l}^{-1}$)	3nm		8nm		11nm		21nm		50nm	
	Graphitic C (Atomic %)	T_{IN} ($^{\circ}\text{C}$) (mean \pm s.e.m.)	Graphitic C (Atomic %)	T_{IN} ($^{\circ}\text{C}$) (mean \pm s.e.m.)	Graphitic C (Atomic %)	T_{IN} ($^{\circ}\text{C}$) (mean \pm s.e.m.)	Graphitic C (Atomic %)	T_{IN} ($^{\circ}\text{C}$) (mean \pm s.e.m.)	Graphitic C (Atomic %)	T_{IN} ($^{\circ}\text{C}$) (mean \pm s.e.m.)
13	60.7	-27.4 ± 0.2	61.1	-27.5 ± 0.1	63.5	-23.8 ± 0.2	66.1	-22.7 ± 0.2	69.4	-21.8 ± 0.1
	65.3	-27.1 ± 0.2	69.7	-25.8 ± 0.2	67.0	-22.6 ± 0.1	68.7	-20.2 ± 0.2	80.1	-17.2 ± 0.2
	68.8	-27.8 ± 0.1	75.3	-25.2 ± 0.2	69.7	-21.7 ± 0.2	71.0	-18.6 ± 0.2	83.9	-14.3 ± 0.2
5.2	60.7	-27.2 ± 0.1	61.5	-27.1 ± 0.2	63.5	-20.0 ± 0.2	66.1	-19.3 ± 0.1	69.4	-17.4 ± 0.2
	65.3	-27.7 ± 0.2	69.7	-24.1 ± 0.1	67.0	-19.3 ± 0.3	68.7	-15.8 ± 0.2	80.1	-13.5 ± 0.2
	68.8	-26.9 ± 0.2	75.3	-21.9 ± 0.2	69.7	-18.7 ± 0.2	71.0	-13.5 ± 0.2	83.9	-12.4 ± 0.2
0.52	60.7	-27.2 ± 0.2	61.5	-27.4 ± 0.2	63.5	-17.9 ± 0.2	66.1	-16.9 ± 0.1	69.4	-15.7 ± 0.1
	65.3	-27.4 ± 0.1	69.7	-23.2 ± 0.2	67.0	-17.6 ± 0.2	68.7	-14.0 ± 0.2	80.1	-11.1 ± 0.1
	68.8	-27.0 ± 0.2	75.3	-20.1 ± 0.2	69.7	-16.9 ± 0.2	71.0	-11.7 ± 0.2	83.9	-10.0 ± 0.2

The volume of the water droplet is 0.2 μl ; cooling rate, 5 $^{\circ}\text{C min}^{-1}$. For each mean, the total number of measurements is about 150. Ice nucleation probability distributions (Gaussian fitting) for each mean T_{IN} are shown in Supplementary Figs. 10, 11 and 12. The content of graphitic C is obtained from the area ratio of the graphitic C sub-peak in the C 1s core-level XPS spectra (Supplementary Fig. 6). The sum of the content of graphitic and oxidative carbon is a unit, so higher content of graphitic carbon represents a lower degree of oxidation.

Earliest hunting scene in prehistoric art

<https://doi.org/10.1038/s41586-019-1806-y>

Received: 9 May 2019

Accepted: 18 October 2019

Published online: 11 December 2019

Maxime Aubert^{1,2,8}, Rustan Lebe³, Adhi Agus Oktaviana^{1,4,8}, Muhammad Tang³, Basran Burhan², Hamrullah⁷, Andi Jusdi³, Abdullah³, Budianto Hakim⁵, Jian-xin Zhao⁶, I. Made Geria⁴, Priyatno Hadi Sulistyarto⁴, Ratno Sardi⁵ & Adam Brumm^{2,8*}

Humans seem to have an adaptive predisposition for inventing, telling and consuming stories¹. Prehistoric cave art provides the most direct insight that we have into the earliest storytelling^{2–5}, in the form of narrative compositions or ‘scenes’^{2,5} that feature clear figurative depictions of sets of figures in spatial proximity to each other, and from which one can infer actions taking place among the figures⁵. The Upper Palaeolithic cave art of Europe hosts the oldest previously known images of humans and animals interacting in recognizable scenes^{2,5}, and of therianthropes^{6,7}—abstract beings that combine qualities of both people and animals, and which arguably communicated narrative fiction of some kind (folklore, religious myths, spiritual beliefs and so on). In this record of creative expression (spanning from about 40 thousand years ago (ka) until the beginning of the Holocene epoch at around 10 ka), scenes in cave art are generally rare and chronologically late (dating to about 21–14 ka)⁷, and clear representations of therianthropes are uncommon⁶—the oldest such image is a carved figurine from Germany of a human with a feline head (dated to about 40–39 ka)⁸. Here we describe an elaborate rock art panel from the limestone cave of Leang Bulu’ Sipong 4 (Sulawesi, Indonesia) that portrays several figures that appear to represent therianthropes hunting wild pigs and dwarf bovids; this painting has been dated to at least 43.9 ka on the basis of uranium-series analysis of overlying speleothems. This hunting scene is—to our knowledge—currently the oldest pictorial record of storytelling and the earliest figurative artwork in the world.

Previous uranium-series (U-series) dating has suggested that the oldest known figurative cave art is found in Indonesia^{9–11}. Up until now, the earliest minimum U-series ages for representative artworks reflect dates of 40 ka for a naturalistic painting of a wild bovid in Kalimantan¹⁰ and, from south Sulawesi, 35.4 ka for a painting of a pig—possibly a female babirusa⁹ or young Sulawesi warty pig (*Sus celebensis*)¹². Non-figurative rock art dated to 65 ka in Spain¹³ has been attributed to Neanderthals, but this claim has been questioned on various grounds^{14–16}. With a minimum age of 40.8 ka, the earliest dated art that is generally attributed to modern humans in Europe is an abstract ‘disc’ sign from the rock art site of El Castillo in Spain¹⁷. Although animal motifs are abundant in the Pleistocene cave art of Indonesia^{9–11} and Europe^{7,17}, in both regions humans hunting fauna are very seldom depicted; composite human–animal figures are also uncommon. In Europe, images of lone animals that are seemingly impaled by projectiles are documented in art of Magdalenian cultures (dating to about 21–14 ka)¹⁸; however, the motifs that are regarded by some as spears or arrows are subject to varying interpretations⁷. In terms of parietal imagery, one of very few obvious narrative compositions⁵ is the famous scene from the shaft (or ‘well’) at Lascaux (France)^{19,20} (Extended Data Fig. 1). This Magdalenian rock art panel apparently depicts a bird-headed man being charged by a wounded bison^{19,20}. The shaft scene is the subject of considerable speculation²⁰, but some scholars believe it represents a real hunt⁷; if this is the case, so far as we can ascertain this would be

the oldest narrative composition that portrays a hunting scene in European art. The earliest image that is generally accepted to represent a therianthrope is the Löwenmensch (‘lion-man’) figurine, a 31.1-cm-tall mammoth-ivory statuette of an apparently part-human, part-lion

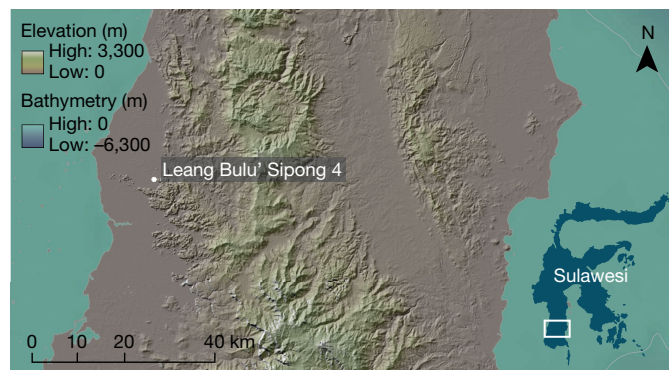


Fig. 1 | Site location in Sulawesi, Indonesia. The limestone cave of Leang Bulu’ Sipong 4 is a rock art site in the tower karst region of Pangkep. Map data: STRM1 Arc-Second Global by NASA/NGS/USGS and GEBCO_2014 Grid version 20150318 (<http://gebco.net>). Base map created by M. Kottermair and A. Jalandoni.

¹Place, Evolution and Rock Art Heritage Unit (PERAHU), Griffith Centre for Social and Cultural Research, Griffith University, Gold Coast, Queensland, Australia. ²Australian Research Centre for Human Evolution, Environmental Futures Research Institute, Griffith University, Brisbane, Queensland, Australia. ³Balai Pelestarian Cagar Budaya, Makassar, Indonesia. ⁴Pusat Penelitian Arkeologi Nasional (ARKENAS), Jakarta, Indonesia. ⁵Balai Arkeologi Sulawesi Selatan, Makassar, Indonesia. ⁶School of Earth and Environmental Natural Sciences, University of Queensland, Brisbane, Queensland, Australia. ⁷Unaffiliated: Hamrullah. ⁸These authors contributed equally: Maxime Aubert, Adhi Agus Oktaviana, Adam Brumm. *e-mail: a.brumm@griffith.edu.au

creature from Hohlenstein–Stadel (Germany)⁸ (Extended Data Fig. 1). This artefact, which belongs to the early Aurignacian tradition (dating to about 40–39 ka), is regarded by some as the earliest evidence for the capacity to link the concepts of ‘animal’ and ‘person’ into a single abstract category⁴. The Hohlenstein–Stadel figure also has a prominent role in scientific debates about the origins of religion³, as it has been argued that the ability to imagine the existence of things that do not exist—including therianthropes—forms the basis for religious thought³. In Kalimantan, U-series dating¹⁰ has shown that people began painting small anthropomorphic figures inside caves at least 14 ka, and perhaps as early as 21–20 ka—these figures are sometimes shown pursuing deer, but dates are not available for any of these scenes¹⁰. To our knowledge, no unambiguous depictions of therianthropes have previously been identified in the early cave art of Kalimantan or Sulawesi.

The oldest rock art on Sulawesi is found in Maros–Pangkep (Fig. 1), an approximately 450-km² area of limestone karst in the south of the island²¹, where excavations have revealed archaeological evidence for human habitation by at least 50–40 ka²². Cave art was first reported in this near-coastal lowland region in the 1950s²³, and at least 242 caves and shelters that contain parietal imagery have thus far been documented^{24–26}; new sites are found each year. In December 2017, we discovered Leang Bulu’ Sipong 4, a rock art site in a high-level limestone cave in Pangkep (Fig. 1, Extended Data Fig. 2). The rear cave wall features a 4.5-m-wide rock art panel with monochrome paintings of what we interpret as therianthropic figures hunting endemic mammals (Fig. 2, Extended Data Figs. 3–6, Supplementary Fig. 1). Concerning the latter, six individual animal motifs are identifiable: two suids (denoted pig 1 and pig 2)—most probably *S. celebensis*—and four dwarf buffaloes (anoas, *Bubalus* sp.)²⁷ (denoted anoa 1 to anoa 4) (Extended Data Figs. 3, 5, 6). Associated with these suids and bovids are at least eight small figures that are human-like in form but have animal characteristics (denoted therianthrope 1 to therianthrope 8), several of which appear to be holding long thin objects that we interpret as spears and/or ropes. There are no indications that these figures were added at a later period of time to a panel that previously contained only animal motifs. Both motif types (animals and therianthropic figures) were produced in broadly the same artistic style associated with early figurative painting in Maros–Pangkep⁹ and using the same technique and dark red pigment; moreover, these figures all exhibit comparable states of weathering, suggesting contemporaneity. Furthermore, the juxtaposition of the figurative animal motifs and therianthropes 1–8 in this rock art panel is in our view indicative of a scene in the modern, Western sense of the term⁵.

Therianthropes 1–8 are simplified and highly stylized forms that in some instances exhibit elongated lower faces, which we regard as resembling muzzles or snouts (for example, therianthrope 1), along with other animal-like morphological characters. This manner of depiction is superficially similar to a stylistic convention that is believed by some to have been used to represent the rare humans in early European cave art²⁸. However, we consider it unlikely that the Leang Bulu’ Sipong 4 figures explicitly depict human beings: at least one of them is apparently portrayed with a tail (therianthrope 1) and another appears to

have a beak (therianthrope 4) (Fig. 2). It is possible that these figures portray human hunters wearing skins, masks or other animal body parts as camouflage⁷; however, this would have the improbable implication that hunters were disguising themselves as small animals such as birds. Because therianthropes 1–8 appear to display characteristics of both people and animals (albeit of indeterminate species), it is reasonable in our estimation to interpret these motifs as intentional representations of therianthropes.

The apparent depiction of therianthropes could imply that some, or all, aspects of this imagery may not pertain to human experiences in the real world⁶. Further support for this is provided by the notably small size of the therianthropic figures in relation to the animals, given that the pig species represented (*S. celebensis*) reaches a maximum body height of only 60 cm¹² and anoas (about 100 cm in height) are the most diminutive taxon of wild cattle²⁷. Therianthropes in prehistoric art are often attributed—though not uncontroversially—to shamanic beliefs and visions, such as representing ‘animal spirit helpers’^{6,29}. Whether such interpretations are appropriate in the case of Leang Bulu’ Sipong 4, or whether the apparent portrayal of therianthropes suggests that the image-makers perceived themselves as an indivisible part of the animal world, is uncertain. Whatever the case, it seems evident that this complex scene, with its multiple interacting subjects, is rich in narrative content.

To date this multi-figured artwork, we conducted U-series analysis on four coralloid speleothems that are directly associated with three of the animal motifs (Table 1, Extended Data Figs. 7–10, Supplementary Table 1). Samples BSP4.2 and BSP4.3 were overlying the hindquarter of pig 1 (Fig. 2, Extended Data Figs. 7, 8). Sample BSP4.4 overlays one of the horns of anoa 2 (Fig. 2, Extended Data Fig. 9) and sample BSP4.5 formed on top of the hindquarter of anoa 3 (Fig. 2, Extended Data Fig. 10). Pig 1, anoa 2 and anoa 3 are associated with therianthrope 1, therianthrope 2, and therianthropes 3–8, respectively. Individual samples were each divided into a series of 5 aliquots (except for sample BSP4.4, which was divided into 3 aliquots), giving a total of 18 U-series age determinations (Table 1, Methods, Supplementary Table 1). The resultant dates for samples BSP4.2 and BSP4.3 are in stratigraphic order or yielded indistinguishable ages within uncertainties, demonstrating closed-system conditions for uranium and thorium. Samples BSP4.4 and BSP4.5 each display a slight age reversal, which we attribute to individual aliquots sampling a series of concentric growth layers and/or ‘mounds’ of widely varying ages (Methods). The oldest minimum U-series age is for pig 1, for which we obtained a date of 43.9 ka (BSP4.3). This is supported by another sample (BSP4.2) over the same figure that has a minimum date of 35.1 ka. Additionally, we obtained minimum dates of 41 and 40.9 ka for anoa 2 (BSP4.4) and anoa 3 (BSP4.5), respectively.

U-series dating demonstrates that the rock art scene at Leang Bulu’ Sipong 4 is the oldest known parietal art created by modern humans. The portrayal of multiple hunters confronting at least two separate prey species possibly suggests a game drive, a communal hunt in which animals are indiscriminately flushed from cover and directed towards waiting hunters—if this is the case, this scene would be the oldest known visual record of a hunting strategy. Of further note is

Table 1 | U-series dating results for coralloid speleothems from Leang Bulu’ Sipong 4

Sample	Description	Sample weight (mg)	²³⁸ U (ppm)	²³⁰ Th/ ²³² Th	²³⁰ Th/ ²³⁸ U	²³⁴ U/ ²³⁸ U	Uncorrected date ± 2σ (ka)	Corrected date ± 2σ (ka)	Corrected initial ²³⁴ U/ ²³⁸ U
BSP4.2.5	Overlies pig 1	3.1	2.4	29	0.2278 ± 0.0014	0.8084 ± 0.0013	36.62 ± 0.28	35.70 ± 0.57	0.7867 ± 0.0017
BSP4.3.5	Overlies pig 1	2.9	2.2	75	0.2744 ± 0.0020	0.8248 ± 0.0017	44.83 ± 0.44	44.41 ± 0.49	0.8008 ± 0.0020
BSP4.4.1	Overlies anoa 2	9.1	2.2	246	0.2970 ± 0.0013	0.9438 ± 0.0010	41.38 ± 0.23	41.26 ± 0.24	0.9368 ± 0.0011
BSP4.5.5	Overlies anoa 3	3.6	2.1	92	0.3060 ± 0.0020	0.9660 ± 0.0018	41.63 ± 0.34	41.32 ± 0.38	0.9617 ± 0.0020

n = 4 coralloid speleothems. The table reports only the oldest minimum age for individual coralloid samples. For full dating results, see Supplementary Table 1. Ratios are activity ratios calculated from the atomic ratios; errors are at 2σ level. The dates are calculated using Isoplot 3.75 Program³⁰, with previously published decay constants³¹. Corrected dates were calculated assuming an initial/detrital ²³⁰Th/²³²Th activity ratio equal to 0.825 (±50%) (the bulk-Earth value, which is most commonly used for initial/detrital ²³⁰Th corrections).

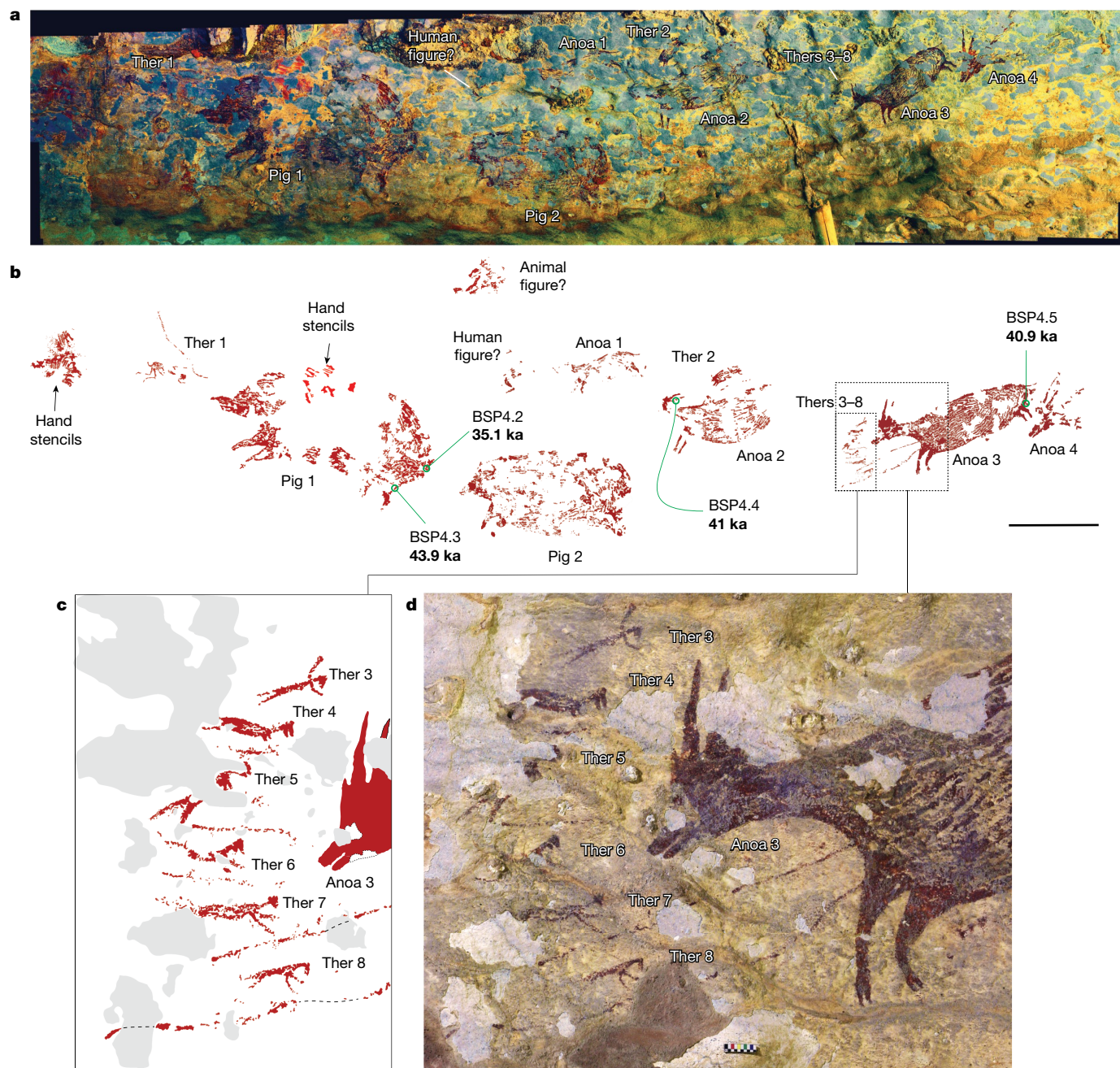


Fig. 2 | Dated rock art panel at Leang Bulu' Sipong 4. **a**, Photostitched panorama of the rock art panel (using photographs enhanced with DStretch). Ther, therianthrope. **b**, Tracing of rock art panel showing results of U-series dating. Scale bar, 20 cm. **c**, **d**, Detail of a group of therianthrope (part-human, part-animal) figures confronting an anoa. As evident in **a**, **c**, **d**, the surface of the cave wall is extensively exfoliated, erasing some of the art. However, the following elements (from left to right; see **a**, **b**) are clearly discernible: a hunter (therianthrope 1; 26 × 12 cm) apparently spearing or roping a pig (pig 1; 123 × 58 cm), in which the body of the hunter appears to be human-like in form but has a tail (Extended Data Figs. 3, 4); a lone pig (pig 2; 84 × 42 cm) (Extended Data Fig. 3); a small anoa (anoa 1; 51 × 24 cm) with an unidentified (possibly human) figure beside it (Extended Data Fig. 5); a small figure (therianthrope 2), the head of which is missing and which is positioned above an anoa (anoa 2; 74 × 29 cm) that it is possibly spearing or roping—the figure appears to be a fully composite being that apparently combines the characteristics of a human and two kinds of non-human animals (an anoa and a reptile) (Extended Data Fig. 5); a group of six very small figures (about 4–8 cm tall) (therianthropes 3–8) confronting an

anoa (anoa 3) with ropes or spears—these tiny figures generally have anthropomorphic bodies, but heads and/or other body parts that are animal-like in form (for close-up images of each of the figures, see Extended Data Fig. 6). Another anoa (anoa 4) is positioned behind anoa 3, but only its head and back line are complete. The locations of four coralloid speleothems (samples BSP4.2 to BSP4.5) collected in association with three animal figures, and which yielded minimum U-series ages for the rock art panel, are indicated in the central tracing (**b**). The earliest minimum date for each sample is provided. The only elements that are evidently not coeval in time with the therianthropes and animals are two separate clusters of hand stencils³²; these motifs were created using a lighter shade of red pigment and are differentially weathered, and one group of stencils was clearly superimposed onto pig 1 following a period of weathering of the cave wall surface, indicating a considerable time lapse between these two phases of art production (Extended Data Fig. 3). In Maros–Pangkep, the oldest dated hand stencil of the distinctive narrow-fingered style³² has a minimum U-series age of 17.8 thousand years⁹.

the possible depiction of ropes in this artwork, which perhaps implies that Late Pleistocene hunters in Sulawesi engaged in the dangerous activity of capturing live adult pigs and anoa. Although the meanings of the imagery are uncertain and likely to remain so, this rock art scene may be regarded not only as the earliest dated figurative art in the world but also as the oldest evidence for the communication of a narrative in Palaeolithic art. This is noteworthy, given that the ability to invent fictional stories may have been the last and most crucial stage in the evolutionary history of human language and the development of modern-like patterns of cognition¹. The figures that we interpret as therianthropes are also the earliest images of this kind yet discovered. These figures are perhaps twice as old as the ‘birdman’ in the much-discussed shaft scene at Lascaux^{19,20}, and at least several millennia older than the iconic lion-headed figurine from Aurignacian Germany⁸. Our findings therefore further suggest that the first known indication of religious-like thinking—the ability to conceive of non-real entities such as therianthropes³—comes not from Europe as has long been assumed^{3,4}, but occurs at least 43.9 ka in Sulawesi. The conspicuousness of therianthropes in the oldest recorded hunting scenes also offers hints at the deeply rooted symbolism of the human–animal bond and predator–prey relationships in the spiritual beliefs, narrative traditions and image-making practices^{1,4} of our species.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1806-y>.

1. Boyd, B. The evolution of stories: from mimesis to language, from fact to fiction. *Wiley Interdiscip. Rev. Cogn. Sci.* **9**, e1444 (2018).
2. Azéma, M. & Rivère, F. Animation in Palaeolithic art: a pre-echo of cinema. *Antiquity* **86**, 316–324 (2012).
3. Mithen, S. in *Becoming Human: Innovation in Prehistoric Material and Spiritual Culture* (eds Renfrew, C. & Morley, I.) 123–134 (Cambridge Univ. Press, 2009).
4. Wynn, T., Coolidge, F. & Bright, M. Hohlenstein-Stadel and the evolution of human conceptual thought. *Camb. Archaeol. J.* **19**, 73–84 (2009).
5. Davidson, I. in *Making Scenes: Global Perspectives on Scenes in Rock Art* (eds Davidson, I. & Nowell, A.) (Berghahn, in the press).
6. Taçon, P. S. C. & Chippindale, C. in *Theoretical Perspectives in Rock Art Research: ACRA: the Alta Conference on Rock Art* (ed. Hølskog, K.) 175–210 (Oslo: Novus forlag: Instituttet for Sammenlignende Kulturforskning, 2001).
7. Bahn, P. G. & Vertut, J. *Journey through the Ice Age* (Weidenfeld & Nicolson, London, 1997).
8. Kind, C.-J., Ebinger-Rist, N., Wolf, S., Beutelspacher, T. & Wehrberger, K. The smile of the Lion Man. Recent excavations in Stadel Cave (Baden-Württemberg, south-western Germany) and the restoration of the famous Upper Palaeolithic figurine. *Quartär* **61**, 129–145 (2014).
9. Aubert, M. et al. Pleistocene cave art from Sulawesi, Indonesia. *Nature* **514**, 223–227 (2014).
10. Aubert, M. et al. Palaeolithic cave art in Borneo. *Nature* **564**, 254–257 (2018).
11. Aubert, M., Brumm, A. & Taçon, P. S. C. The timing and nature of human colonization of Southeast Asia in the Late Pleistocene – a rock art perspective. *Curr. Anthropol.* **58**, S553–S566 (2017).
12. Burton, J. A., Mustari, A. H. & Rejeki, I. S. in *Ecology, Conservation and Management of Wild Pigs and Peccaries* (eds Melletti M. & Meijaard, E.) 184–192 (Cambridge Univ. Press, 2018).
13. Hoffmann, D. L. et al. U–Th dating of carbonate crusts reveals Neandertal origin of Iberian cave art. *Science* **359**, 912–915 (2018).
14. Aubert, M., Brumm, A. & Huntley, J. Early dates for ‘Neanderthal cave art’ may be wrong. *J. Hum. Evol.* **125**, 215–217 (2018).
15. Pearce, D. G. & Bonneau, A. Trouble on the dating scene. *Nat. Ecol. Evol.* **2**, 925–926 (2018).
16. Slimak, L. et al. Comment on “U–Th dating of carbonate crusts reveals Neandertal origin of Iberian cave art”. *Science* **361**, eaau1371 (2018).
17. Pike, A. W. G. et al. U-series dating of Paleolithic art in 11 caves in Spain. *Science* **336**, 1409–1413 (2012).
18. Allain, J. & Rigaud, A. Les petites pointes dans l’industrie osseuse de La Garenne: fonction et figurations. *Anthropologie* **96**, 135–162 (1992).
19. Davenport, D. & Jochim, M. A. The scene in the shaft at Lascaux. *Antiquity* **62**, 558–562 (1988).
20. Le Quellec, J.-L. *L’homme de Lascaux et l’énigme du Puits* (Camille Bercot, Tautem, 2017).
21. McDonald, R. C. Limestone morphology in South Sulawesi, Indonesia. *Z. Geomorphol.* **26** (Suppl.), 79–91 (1976).
22. Brumm, A. et al. Early human symbolic behavior in the Late Pleistocene of Wallacea. *Proc. Natl Acad. Sci. USA* **114**, 4105–4110 (2017).
23. van Heekeren, H. R. Rock-paintings and other prehistoric discoveries near Maros (South West Celebes). *Laporan Tahunan Dinas Purbakala* **1950**, 22–35 (1952).
24. Eriawati, Y. *Lukisan di Gua-Gua Karst Maros–Pangkep, Sulawesi Selatan: Gambaran Penghuni dan Matapencahariannya* (Indonesian Ministry of Cultural Media Development, 2003).
25. Kurniawan, R. et al. Chemistry of prehistoric rock art pigments from the Indonesian island of Sulawesi. *Microchem. J.* **146**, 227–233 (2019).
26. Saiful, A. M. & Burhan, B. Lukisan fauna, pola sebaran dan lanskap budaya di kawasan kars Sulawesi bagian selatan. *Walennae* **15**, 75–88 (2017).
27. Groves, C. P. & Grubb, P. *Ungulate Taxonomy* (John Hopkins Univ. Press, 2011).
28. Lorblanchet, M. in *Animals into Art* (ed. Morphy, H.) 109–141 (Unwin Hyman, London, 1989).
29. Lewis-Williams, D. J. *The Mind in the Cave: Consciousness and the Origins of Art* (Thames & Hudson, 2002).
30. Ludwig, K. R. *User’s Manual for Isoplot 3.75. A Geochronological Toolkit for Microsoft Excel* (Berkeley Geochronology Center Special Publication No. 5) (Berkeley Geochronology Center, Berkeley, 2012).
31. Cheng, H. et al. The half-lives of uranium-234 and thorium-230. *Chem. Geol.* **169**, 17–33 (2000).
32. Oktaviana, A. A. et al. Hand stencils with and without narrowed fingers at two new rock art sites in Sulawesi, Indonesia. *Rock Art Res.* **33**, 32–48 (2016).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

U-series dating

A small segment (about 25–150 mm²) of each coralloid speleothem ($n = 4$) was removed from the rock art panel at Leang Bulu' Sipong 4 using a battery-operated rotary tool equipped with a diamond saw blade. Each speleothem sample was sawn *in situ* so as to produce a continuous microstratigraphic profile that extends from the outer surface of the speleothem through the pigment layer and into the underlying rock face. In the laboratory, the samples were micro-excavated in arbitrary 'spits' over the entire surface of each speleothem, creating a series of five aliquots per sample (except for sample BSP4, which has three aliquots) measuring less than 1 mm in thickness. The red pigment layer corresponding to the artwork was visible across the entire length of each sample. In total, we obtained 18 U-series age determinations (Table 1 and Supplementary Table 1).

The coralloid speleothem samples collected in this study formed from thin films of water on cave surfaces over a long period of time. When precipitated from saturated solutions and under ideal conditions, calcium carbonate usually contains small amounts of soluble uranium (²³⁸U and ²³⁴U), which eventually decay to ²³⁰Th. The latter is essentially insoluble in cave waters and will not precipitate with the calcium carbonate. This produces disequilibrium in the decay chain, in which not all isotopes in the series are decaying at the same rate. Subsequently, ²³⁸U and ²³⁴U decay to ²³⁰Th until secular equilibrium is reached. Because the decay rates are known, the precise measurement of these isotopes enables the calculation of the age of the carbonate formation³³.

U-series dating was carried out using a Nu Plasma multi-collector inductively-coupled plasma mass spectrometer (MC-ICP-MS) in the Radiogenic Isotope Facility at the School of Earth and Environmental Sciences, University of Queensland, following chemical treatment procedures and MC-ICP-MS analytical protocols that have previously been described^{34–36}. Powdered subsamples weighing 3–13 mg were spiked with a mixed ²²⁹Th–²³³U tracer and then completely dissolved in concentrated HNO₃. After digestion, each sample was treated with H₂O₂ to decompose trace amounts of organic matter (if any) and to facilitate complete sample–tracer homogenization. Uranium and thorium were separated using conventional anion-exchange column chemistry using Bio-Rad AG 1-X8 resin. After stripping off the matrix from the column using double-distilled 7 N HNO₃ as eluent, 3 ml of a 2% HNO₃ solution mixed with trace amount of HF was used to elute both uranium and thorium into a 3.5-ml pre-cleaned test tube, ready for MC-ICP-MS analyses, without the need for further drying down and re-mixing. After column chemistry, the U–Th mixed solution was injected into the MC-ICP-MS through a DSN-100 desolvation nebulizer system with an uptake rate of around 0.06 ml min^{−1}. U–Th isotopic ratio measurement was performed on the MC-ICP-MS using a detector configuration to allow simultaneous measurements of both uranium and thorium^{36,37}. The ²³⁰Th/²³⁸U and ²³⁴U/²³⁸U activity ratios of the samples were calculated using previously published decay constants³¹. In some instances, there was enough solution leftover to run the sample another time, producing a 'repeat' (Table 1). U–Th dates were calculated using the Isoplot/Ex 3.75 Program³⁰. In the text, minimum dates are quoted as measured age minus 2σ, rounded to one decimal place.

It is common for secondary calcium carbonate to be contaminated by detrital materials, such as wind-blown or waterborne sediments, a process that can lead to U-series ages that are erroneously older than the true age of the sample. This is due to the pre-existing ²³⁰Th present in the detrital components, which is in some ways analogous to the radiocarbon marine reservoir effect. As the detrital/initial ²³⁰Th cannot

be physically separated from the radiogenic ²³⁰Th for measurement, its contribution to the calculated ²³⁰Th age of the sample is often corrected for using an assumed ²³⁰Th/²³²Th activity ratio in the detrital component. Given the detrital component within a cave is often composed of wind-blown or waterborne sediments that chemically approach the average continental crust, the mean bulk-Earth or upper continental crustal value of ²³²Th/²³⁸U = 3.8, corresponding to an ²³⁰Th/²³²Th activity ratio of 0.825—with an arbitrarily assigned uncertainty of 50%—has commonly been assumed for detrital/initial ²³⁰Th corrections³⁴. In this regard, the degree of detrital contamination may be reflected by the measured ²³⁰Th/²³²Th activity ratio in a sample, with a higher value (such as >20) indicating a relatively small or insignificant effect on the calculated age and a lower value (<20) indicating that the correction on the age will be considerable³³. Because ²³²Th in the sample is largely present in the detrital fraction and plays no part in the decay chain of uranium, the detrital ²³⁰Th in a sample with a measured ²³⁰Th/²³²Th activity ratio >20 would make up only <0.825/20 (about 4.1% of the total ²³⁰Th in the sample).

Sometimes, the assumed ²³⁰Th/²³²Th activity ratio of 0.825 (±50%) for the detrital component may not cover all situations. If the actual ²³⁰Th/²³²Th activity ratio in the detrital component substantially deviates from this assumed range, the detrital correction scheme may introduce considerable bias, especially to samples with a ²³⁰Th/²³²Th activity ratio <20. In such situations, the ²³⁰Th/²³²Th activity ratio in the detrital component can be obtained through direct measurement of sediments associated with speleothems^{13,38,39}, or computed using isochron methods or stratigraphic constraints⁴⁰. In our case, our samples were relatively pure: the ²³⁰Th/²³²Th activity ratio of individual aliquots ranged from 29 to 369. Corrections for detrital components were therefore calculated assuming the bulk-Earth values.

A conceivable problem with the U-series dating method is that calcium carbonate accretions can behave as an open system for uranium, in which the element can be leached out of the accretions or remobilized⁴¹. In such instances, the calculated ages will be too old because the dating method relies on the accurate measurement of uranium versus its decay product ²³⁰Th. In this study, this problem was tackled by avoiding porous samples and by measuring five aliquots from every sample (except for sample BSP4, which had three aliquots). The ages of these subsamples were mostly in chronological order, confirming the integrity of the dated coralloids. If uranium had leached out of the samples, a reverse age profile would have been evident (the ages would have gotten older towards the surface). Samples BSP4 and BSP5 each display a single data point with a slightly younger age (on the order of a few hundred years) (Table 1). We attribute these inversions to the calcite deposits having accumulated in ring-like formations rather than roughly flat and parallel to the pigment layer (as in typical flowstones)⁴². Alternatively, some coralloid speleothems have complex internal morphologies that reflect their origin as aggregates of a cluster of cylindrical, mound-like calcite structures⁴², leaving overhanging features with gaps between older material that are infilled by carbonate materials of younger age. Because the microsampling procedure involves collecting material from an arbitrary depth above the pigment layer, as opposed to sampling individual laminae, the resultant U-series age could—in some instances—be an average of the older mound material and the younger infill. Whichever is the case, it would not be possible to estimate the proportion of each carbonate component that contributes to the calculated age. In summary, owing to the small size of the dated coralloids BSP4 and BSP5, and the arbitrary nature of the micro-excavated spits, it is likely that individual aliquots average out a series of concentric growth layers and/or mounds of widely varying ages, thus explaining the presence of minor outliers in the dating sequence. Because the individual age estimates presented in this study all represent an average of multiple layers of varying ages, the true minimum age of the underlying artwork is possibly older than that reported here.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All relevant data are available from the corresponding author upon reasonable request.

33. Bourdon, B., Henderson, G. M., Lundstrom, C. C. & Turner, S. P. *Uranium-series Geochemistry* (Mineralogical Society of America, Chantilly, 2003).
34. Zhao, J. X., Yu, K. F. & Feng, Y. X. High-precision ^{238}U – ^{234}U – ^{230}Th disequilibrium dating of the recent past – a review. *Quat. Geochronol.* **4**, 423–433 (2009).
35. Clark, T. R. et al. Spatial variability of initial $^{230}\text{Th}/^{232}\text{Th}$ in modern *Porites* from the inshore region of the Great Barrier Reef. *Geochim. Cosmochim. Acta* **78**, 99–118 (2012).
36. Clark, T. R. et al. Discerning the timing and cause of historical mortality events in modern *Porites* from the Great Barrier Reef. *Geochim. Cosmochim. Acta* **138**, 57–80 (2014).
37. Zhou, H. Y., Zhao, J. X., Wang, Q., Feng, Y. X. & Tang, J. Speleothem-derived Asian summer monsoon variations in Central China during 54–46 ka. *J. Quat. Sci.* **26**, 781–790 (2011).
38. Westaway, K. E. et al. An early modern human presence in Sumatra 73,000–63,000 years ago. *Nature* **548**, 322–325 (2017).
39. St Pierre, E., Zhao, J. X. & Reed, E. Expanding the utility of uranium-series dating of speleothems for archaeological and palaeontological applications. *J. Archaeol. Sci.* **36**, 1416–1423 (2009).
40. Hellstrom, J. U–Th dating of speleothems with high initial ^{230}Th using stratigraphical constraint. *Quat. Geochronol.* **1**, 289–295 (2006).
41. Plagnes, V. et al. Cross dating (Th/U – ^{14}C) of calcite covering prehistoric paintings in Borneo. *Quat. Res.* **60**, 172–179 (2003).

42. Vanghi, V., Frisia, S. & Borsato, A. Genesis and microstratigraphy of calcite coralloids analysed by high resolution imaging and petrography. *Sedim. Geol.* **359**, 16–28 (2017).

Acknowledgements This research was funded by Australian Research Council (ARC) fellowships awarded to M.A. (FT170100025) and A.B. (FT160100119), with further financial support from Griffith University. We thank Indonesia's State Ministry of Research and Technology (RISTEK), I. Mahmud (Balai Arkeologi Sulawesi Selatan) and L. Aksa (Balai Pelestarian Cagar Budaya Makassar) for authorizing the research; and P. T. Semen Tonasa for providing access to the site. We acknowledge M. Kottermair, A. Jalandoni, D. P. McGahan, K. Newman and M. Langley for assistance with figure production. We thank P. Veth, B. David and P. S. C. Taçon for comments on the paper.

Author contributions A.B. and M.A. conceived and led the research with senior collaborators B.H., P.H.S., I.M.G. and R.L. The rock art site was discovered by H. as part of a BPCB Makassar field survey led by M.T., and involving specialist input from A.J. and A. Rock art was recorded and analysed in the field by A.A.O., B.B. and R.S., and A.A.O. produced the digital tracings of parietal motifs. M.A. identified and collected the coralloid speleothem samples at the rock art site and conducted the micromilling and subsampling of each speleothem sample. All in-field sampling involving rock art carried out by M.A. was done under the direct supervision of R.L. J.-x.Z. conducted the U-series dating. M.A. and J.-x.Z. analysed and interpreted U-series data, and discussed and approved correction factors and other methodological details pertinent to the dating results. A.B. and M.A. wrote the paper, with key contributions from the other authors. The figures were produced and/or designed by A.B. M.A. and J.-x.Z. prepared the Methods. All authors reviewed and edited the paper.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1806-y>.

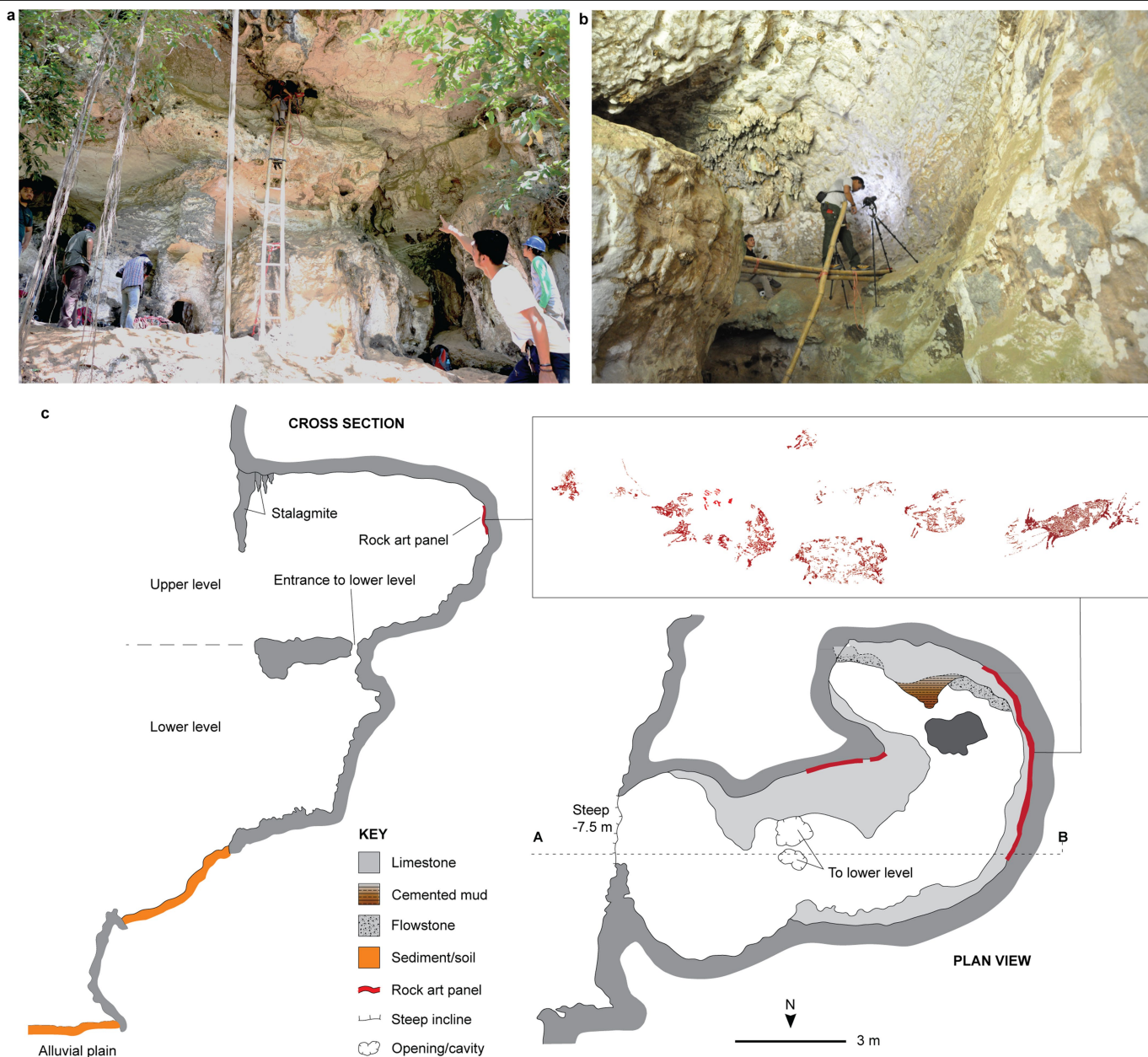
Correspondence and requests for materials should be addressed to A.B.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



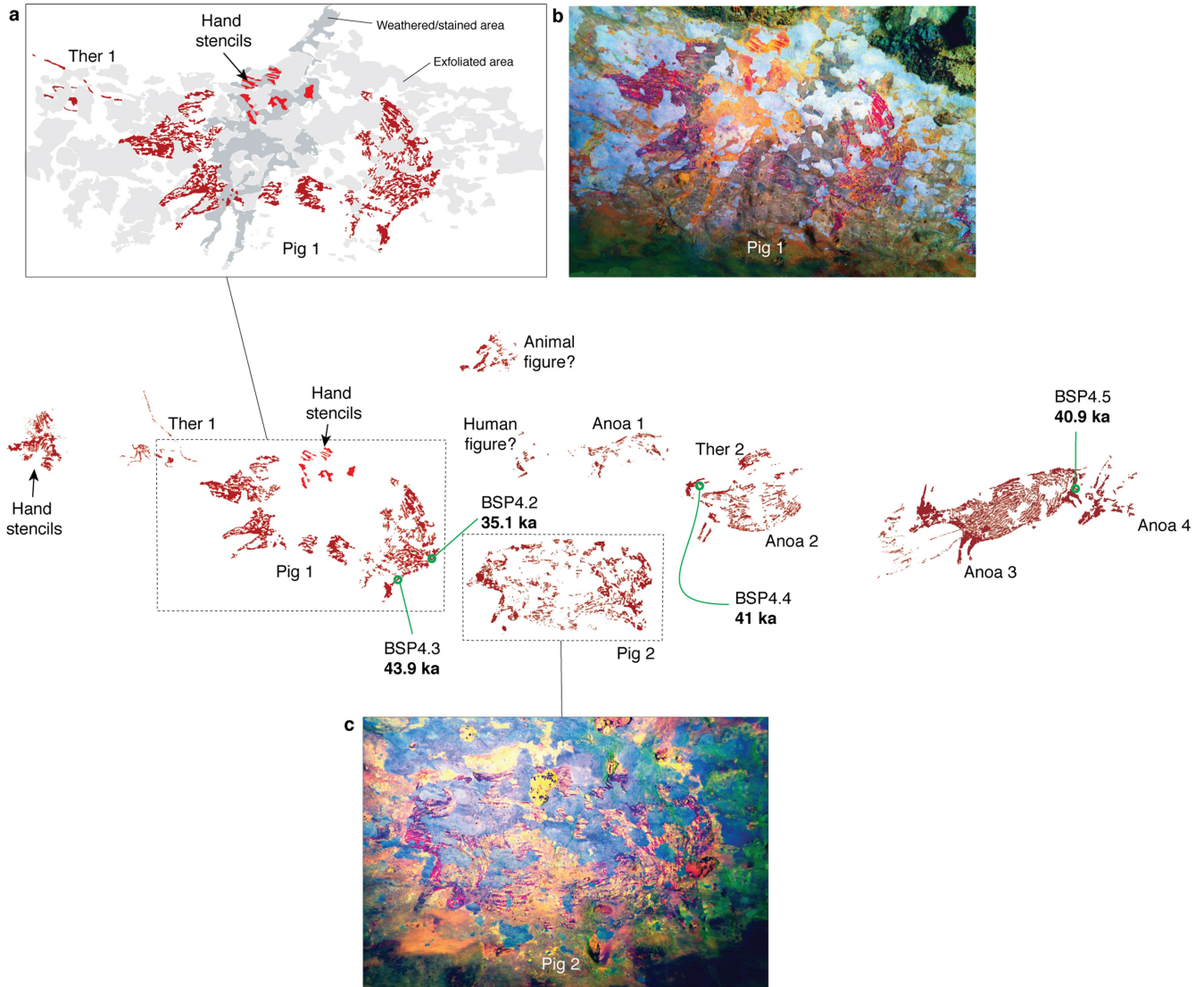
Extended Data Fig. 1 | The oldest hunting scene and therianthrope images known from Europe. **a, b,** The shaft scene from Lascaux (about 21–14 ka) (**a**). This rock art panel is widely interpreted as depicting a bird-headed human figure (**b**) being charged by a bison that it has wounded with a spear; in **a**, the latter object is visible below the partly disembowelled bison. Another object depicted in this scene possibly represents a spearthrower with a sculpted

representation of a bird at the proximal end^{19,20}. **c, d,** The lion-man figurine from Hohlenstein-Stadel⁸. Carved in mammoth ivory, this 31.1-cm-tall image of Aurignacian age (about 40–39 ka) appears to represent a male human figure with the head of a cave lion⁸. The image in **b** is a digital tracing of the relevant section in **a**. Sources: Alamy, used under licence (**a, c**); Shutterstock, used under licence (**d**).



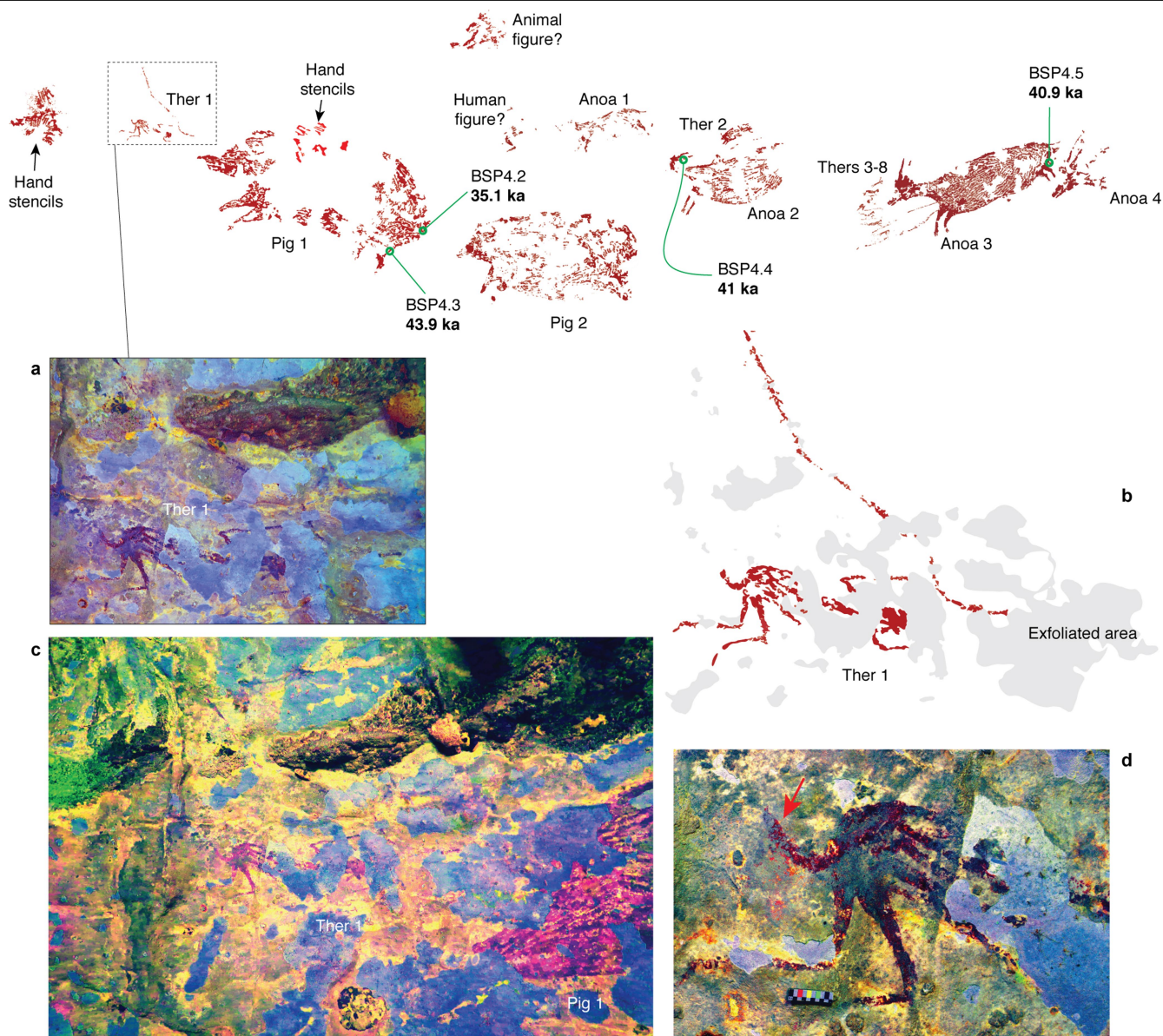
Extended Data Fig. 2 | Leang Bulu' Sipong 4 rock art site. **a, b,** The site is located on the east side of an isolated limestone karst tower. **c,** Cross-section and plan view of the cave site. The cave with the dated rock art panel is positioned in a limestone cliff face and forms the upper level or 'annex', above a valley-floor entrance cave and shelter complex (**a, c**). The entrance to Leang Bulu' Sipong 4 is a small opening about 7.5 m above the ground floor of the lower cave (**a**). The cave is lit by a natural opening on the northeast face (**c**).

The cave itself is formed in a sharply curved phreatic passage measuring 4 m in maximum width, and which is 5.9 m high at the entrance and 5.6 m high at the deepest point inside. The main rock art panel is situated in the light zone on the western wall of the cave, about 3 m above the ground floor surface (**b**). Other rock art inside the cave includes poorly preserved hand stencils and animal paintings. Aside from art, no other evidence for human occupation was observed in the cave.



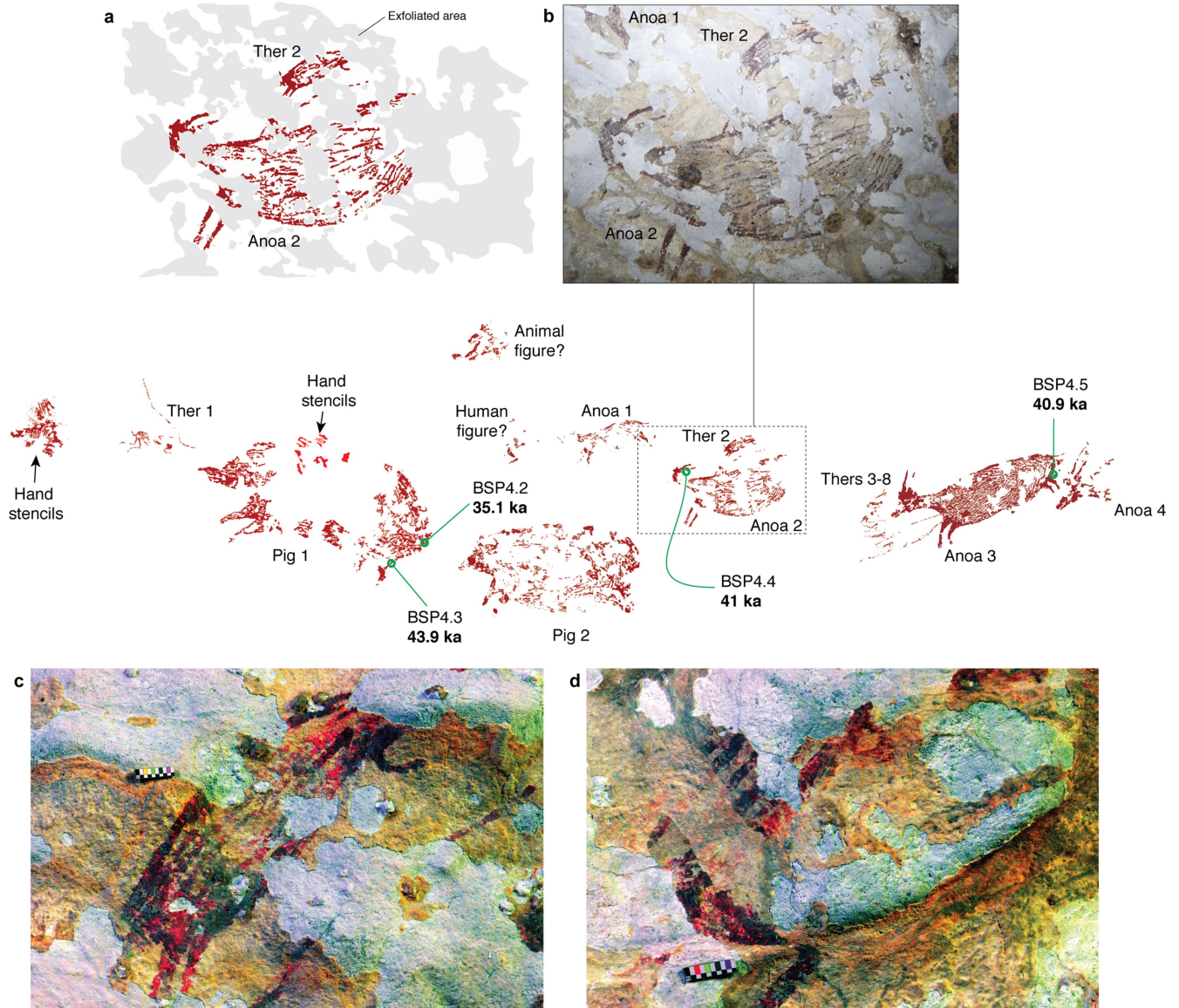
Extended Data Fig. 3 | Details of pig 1 and pig 2. **a, b,** Pig 1 shown in a digital tracing (**a**) and a photograph enhanced using DStretch (**b**). **c,** Photograph of pig 2 enhanced using DStretch. Pig 1 measures 123 × 58 cm. The painting is badly weathered. Much of the body area, and some of the head and mouth, are missing owing to at least two temporally distinct phases of erosion and flaking of the cave-wall surface. In the time that separated these periods of weathering, three narrow-fingered hand stencils³² were created in the upper body area of the pig. No canine tusks are evident, but the animal is apparently portrayed with a row of premolars and molars in the maxilla and mandible; the teeth are

sharp and thus possibly relatively unworn—perhaps indicating that the pig was a relatively young adult. No sexual characteristics are evident. Pig 2 measures 84 × 42 cm and is also substantially deteriorated: most of the head area, and considerable portions of the body, are missing. This pig is positioned to the rear of pig 1 and faces in the same direction as this larger suid. It appears as though it is following behind it. A prominent crest or tuft of head hair, represented by a row of short vertical lines on the crown, is evident in the surviving part of the head area; this is a diagnostic morphological trait of the endemic Sulawesi warty pig (*S. celebensis*)¹².



Extended Data Fig. 4 | Details of therianthrope 1. a, b, Therianthrope 1 shown in a photograph enhanced using DStretch (**a**) and in a digital tracing (**b**). **c,** Photograph of therianthrope 1, enhanced using DStretch, positioned adjacent to the head area of pig 1. On the leftmost side of the panel, therianthrope 1 (26 × 12 cm) is facing towards pig 1 and is possibly crouched down in an active position. In its left hand it is holding a long spear or rope that

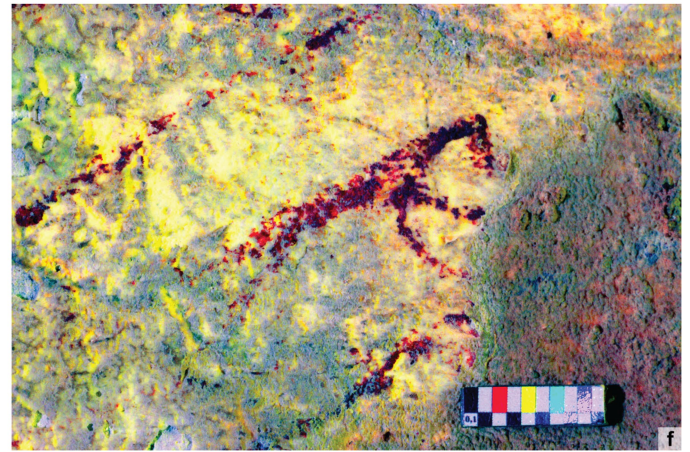
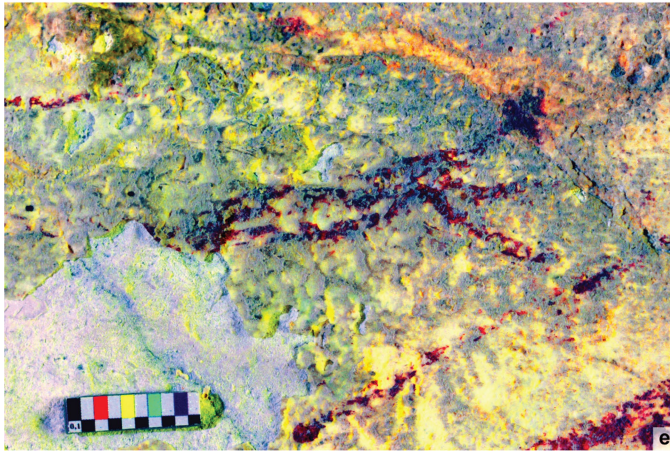
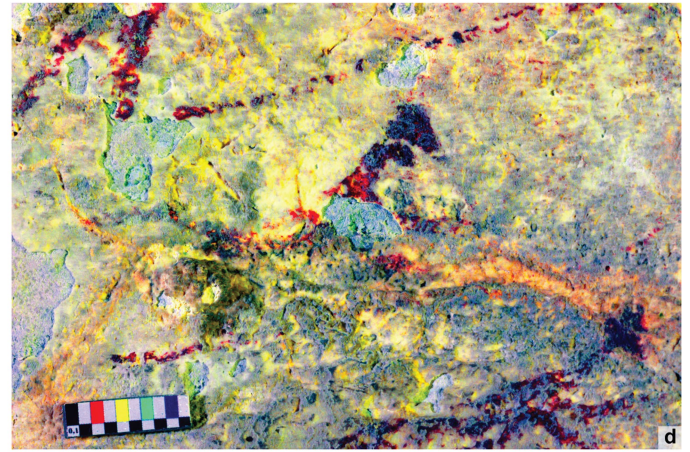
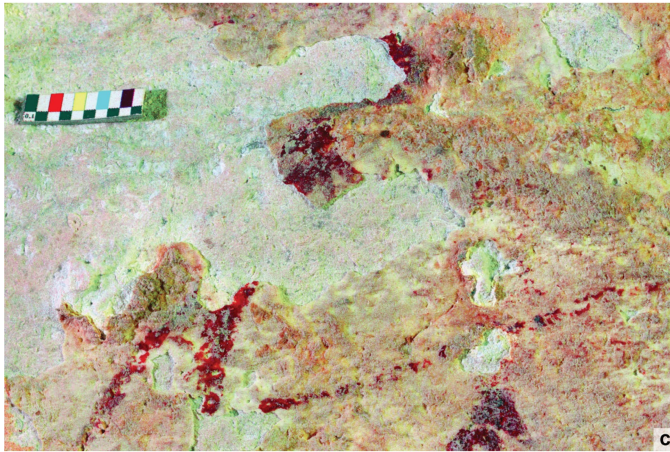
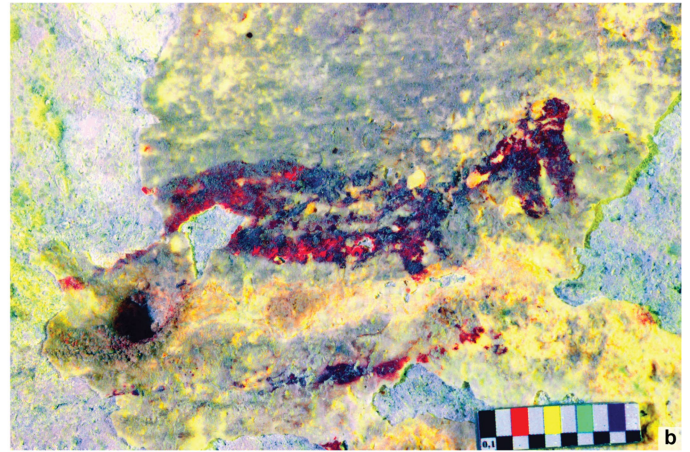
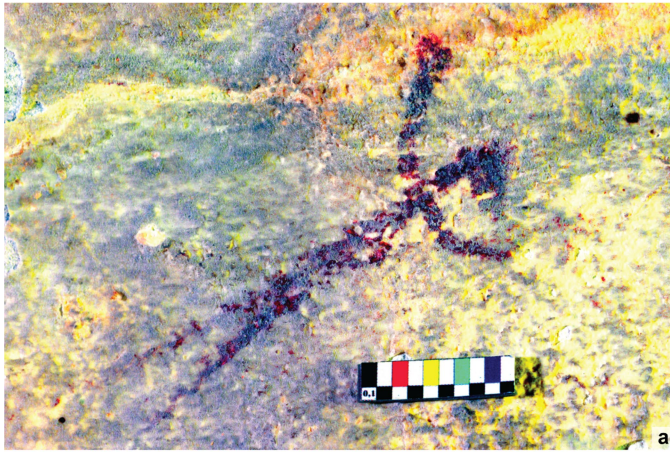
appears to be pointed directly at the head area of this animal, and may once have connected with it; it is not possible to be certain because this part of the panel is missing owing to exfoliation. Therianthrope 1 is depicted with a short, curved mammal-like tail (**d**, highlighted with red arrow). Although the head area of therianthrope 1 is incomplete because of the deterioration of the cave wall, a muzzle or beak-like face is also evident.



Extended Data Fig. 5 | Details of therianthrope 2, anoa 1 and anoa 2.

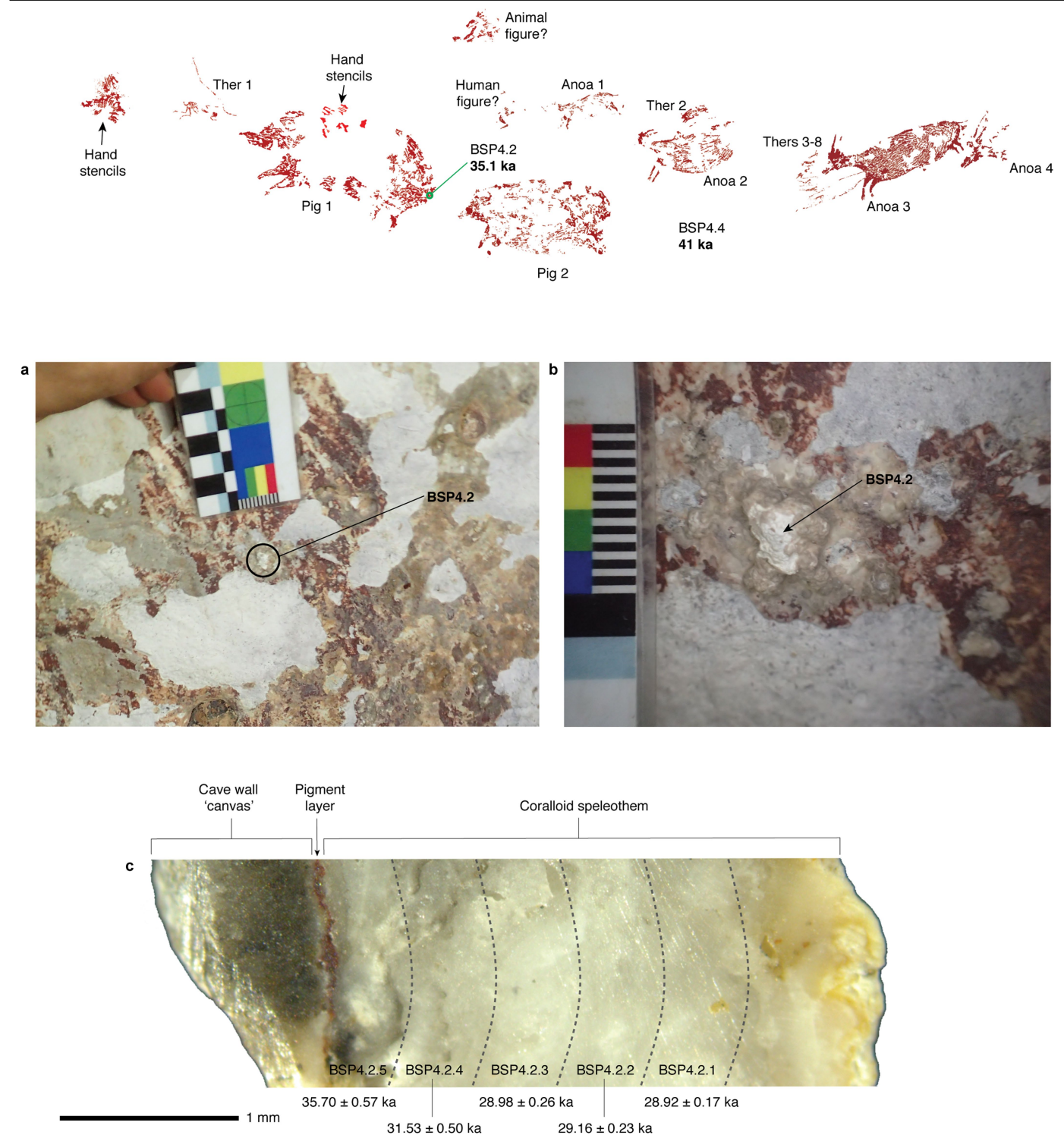
a, b, Therianthrope 2 and anoa 2 shown in a digital tracing (**a**) and photograph (**b**). **c,** Photograph of therianthrope 2, enhancing using DStretch. **d,** Photograph, enhancing using DStretch, of the unidentified, possible human figure to the left of anoa 1. Anoa 2 measures 74 × 29 cm. Although deteriorated, anoa 2 is clearly a dwarf bovid based on the overall body form, long tapering neck and the two straight horns visible in the head area. Therianthrope 2 is much smaller in size than anoa 2, and is positioned directly above it; therianthrope 2 appears to be holding a spear or rope that is entering (or attached to) the back of anoa 2. The area in which the head of therianthrope 2 would have been has been obliterated by exfoliation of the cave-wall surface, but although both of its arms are definitely human-like and it is evidently

grasping a spear or rope, the line of the back and the shape of the neck seem to be notably similar to that of an anoa. Moreover, the bottom half of the figure is distinct from that of the top half, with a tapering profile that possibly merges into the base of a thick tail and with short, curved limbs splayed out to the side. In our opinion, this part of the body resembles the lower half of a lizard or crocodile. It is thus possible that therianthrope 2 represents a composite of at least three different kinds of animals: a human, an anoa and a quadrupedal reptile. Anoa 1, a small and incomplete animal figure (51 × 24 cm) is also visible in this part of the rock art panel. The head is missing but the overall form of the surviving portions of the body (which includes a tail) implies that it is an anoa. A possible human figure adjacent to anoa 1, and another motif above and to the left of it, are too poorly preserved for identification.



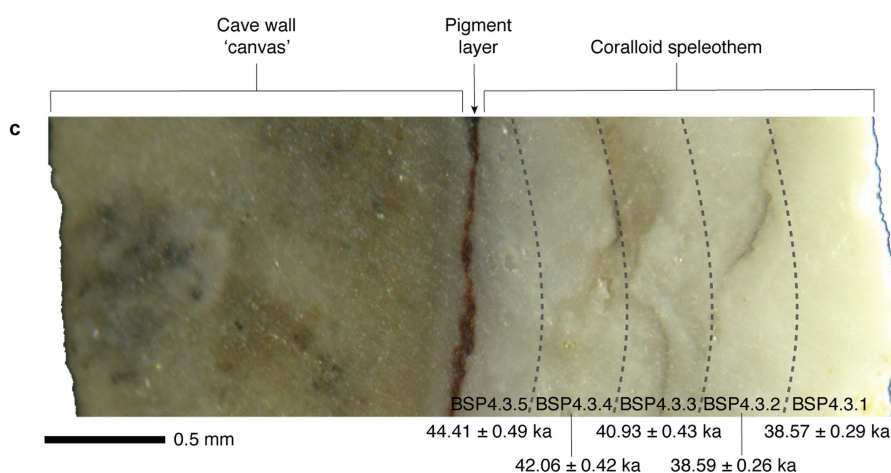
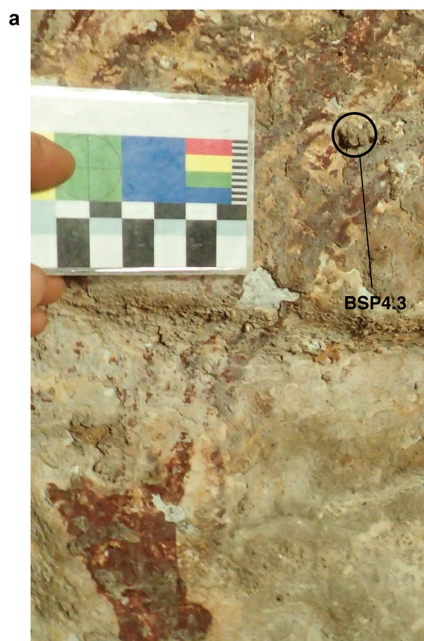
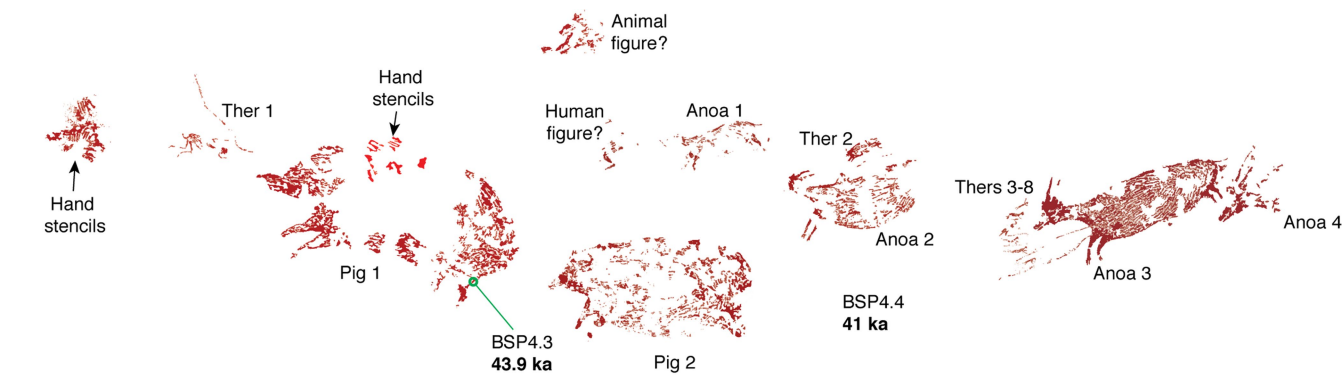
Extended Data Fig. 6 | Details of therianthropes 3–8. All photographs have been enhanced using DStretch. **a**, Therianthrope 3 (5×3 cm) is a stick-like figure with upraised arms and a projecting muzzle-like face. Therianthrope 3 is the only one of the therianthropes at Leang Bulu' Sipong 4 not depicted with a spear or rope. **b**, Therianthrope 4 (6.5×1 cm) is an apparently bird-headed human figure holding a spear or rope. **c**, Therianthrope 5 (8×2 cm) is poorly preserved, but seems to be a human figure with a face similar to that of therianthrope 1. The figure is positioned near an object that may be a spear or rope. **d**, Therianthrope 6 (5×1 cm) has a sinuous reptilian body and a bird-like face. A spear or rope is lying below this figure. **e**, Therianthrope 7 (6×2 cm)

apparently has a human body and upper arms (the legs are too poorly preserved for analysis), but has a pointy head and face that are not human-like in form. This figure is seemingly holding a very long spear or rope that is trailing from the chest area of anoa 3, just below the throat (Fig. 2c, d). **f**, Therianthrope 8 (4×1.7 cm) is also grasping an extremely long spear or rope using two human-like arms, but the shape of the body, neck and head of this figure—especially the elongate, projecting face—are not human-like. The object held by therianthrope 8 appears to trail from the lower neck or upper shoulder of anoa 3 (Fig. 2c, d).



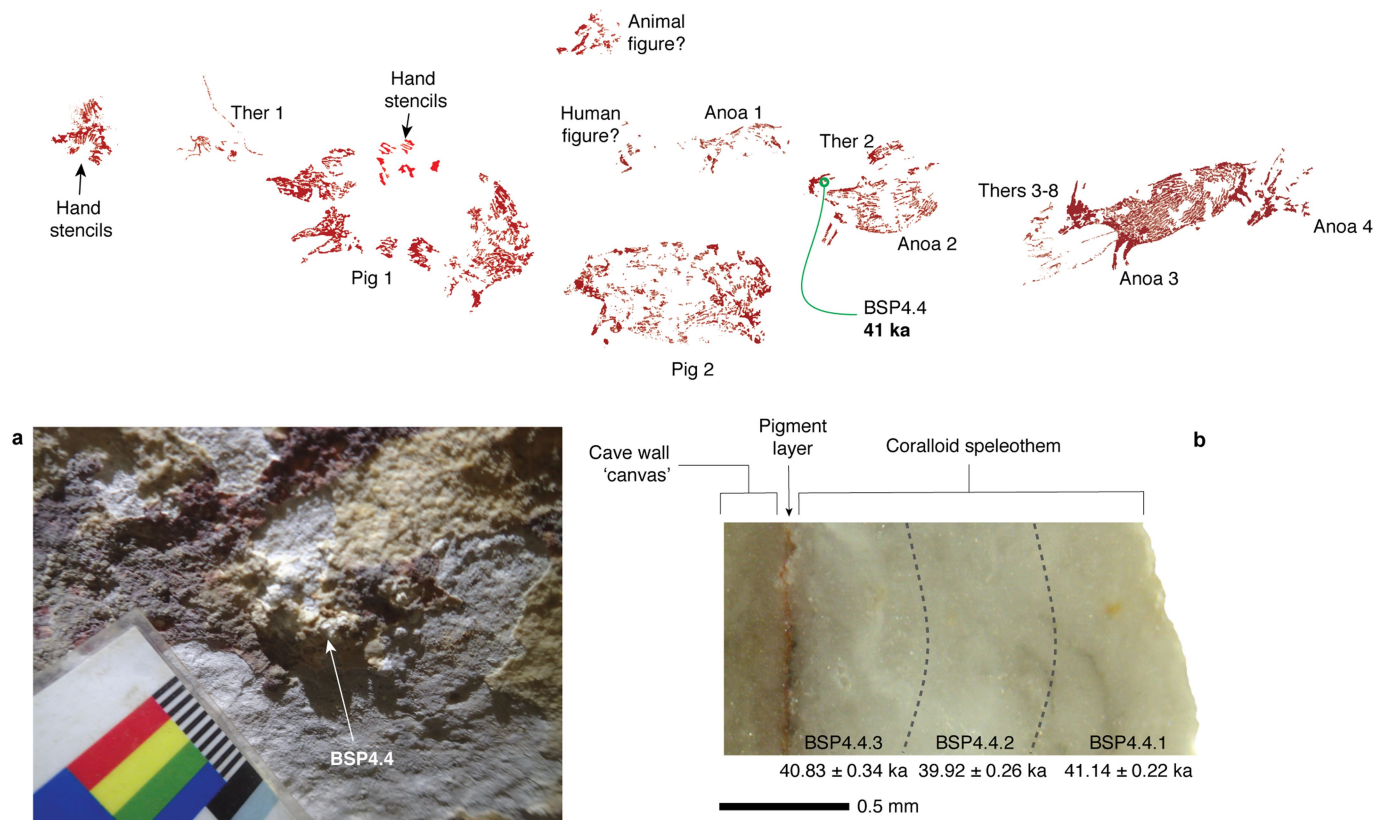
Extended Data Fig. 7 | Coralloid speleothem sample BSP4.2, and U-series dates. **a, b,** Location of the in situ speleothem overlying part of pig 1. **c,** Cross-section of BSP4.2 showing the pigment layer sandwiched between the cave-wall surface and layers of calcite comprising the speleothem that formed over the artwork. Solution U-series dates for a total of five micromilled subsamples

($n = 5$) (BSP4.2.1 to BSP4.2.5) are indicated. The dotted lines represent schematically the micromilling spits used during the subsampling procedure. Minimum dates are quoted as the measured age minus 2σ , rounded to two decimal places.



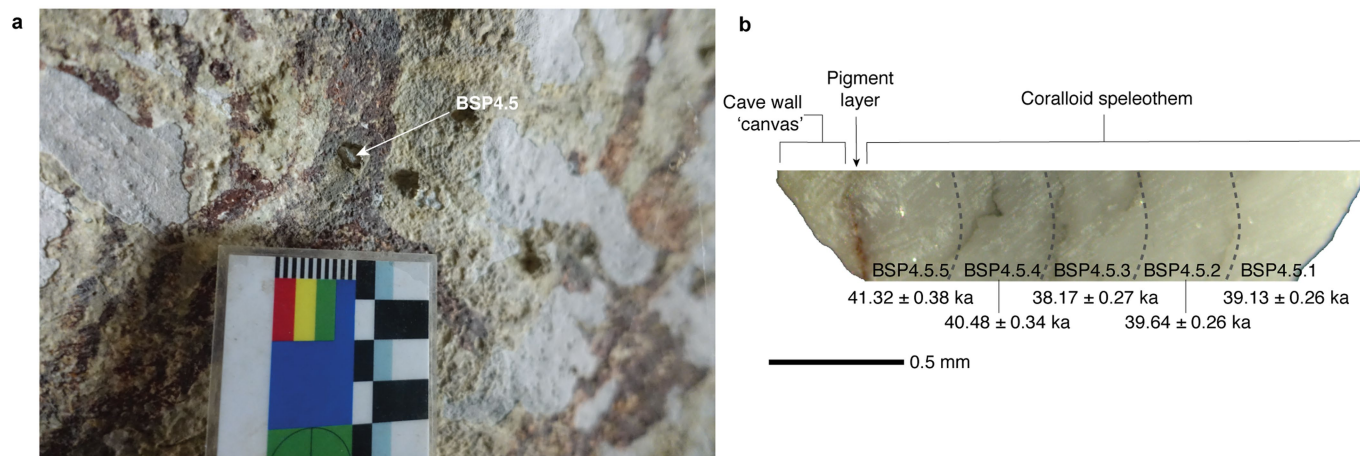
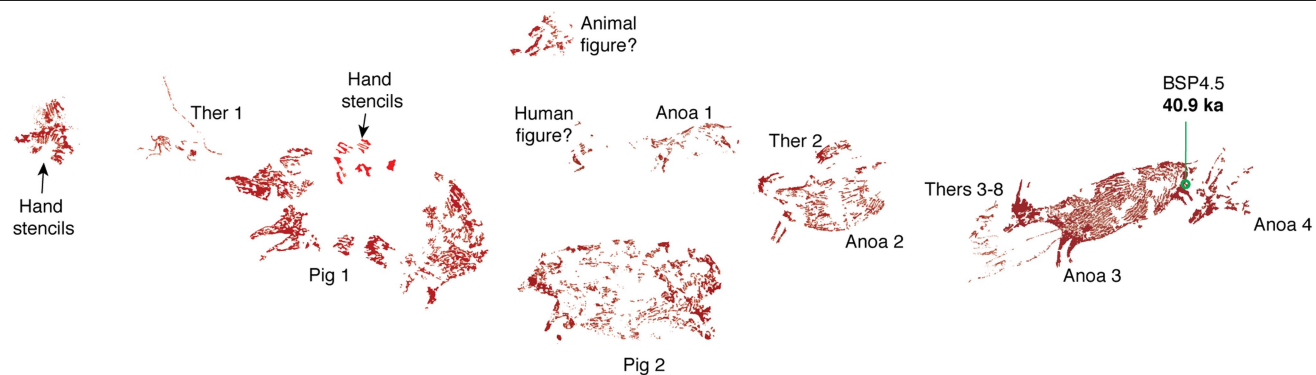
Extended Data Fig. 8 | Coralloid speleothem sample BSP4.3, and U-series dates. **a, b**, Location of the in situ speleothem overlying part of pig1. **c**, Cross-section of BSP4.3 showing the pigment layer sandwiched between the cave-wall surface and layers of calcite comprising the speleothem that formed over the artwork. Solution U-series dates for a total of five micromilled subsamples

($n = 5$) (BSP4.3.1 to BSP4.3.5) are indicated. The dotted lines represent schematically the micromilling spits used during the subsampling procedure. Minimum dates are quoted as measured age minus 2σ , rounded to two decimal places.



Extended Data Fig. 9 | Coralloid speleothem sample BSP4.4, and U-series dates. **a**, Location of the in situ speleothem overlying part of anoa 2. **b**, Cross-section of BSP4.4 showing the pigment layer sandwiched between the cave-wall surface and layers of calcite comprising the speleothem that formed over the artwork. Solution U-series dates for a total of three ($n = 3$) micromilled

subsamples (BSP4.4.1 to BSP4.4.3) are indicated. The dotted lines represent schematically the micromilling spits used during the subsampling procedure. Minimum dates are quoted as measured age minus 2σ , rounded to two decimal places.



Extended Data Fig. 10 | Coraloid speleothem sample BSP4.5, and U-series dates. **a**, Location of the in situ speleothem overlying part of anoa 3. **b**, Cross-section of BSP4.5 showing the pigment layer sandwiched between the cave-wall surface and layers of calcite comprising the speleothem that formed over the artwork. Solution U-series dates for a total of five ($n=5$) micromilled

subsamples (BSP4.5.1 to BSP4.5.5) are indicated. The dotted lines represent schematically the micromilling spits used during the subsampling procedure. Minimum dates are quoted as measured age minus 2σ , rounded to two decimal places.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|---|
| n/a | Confirmed |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

n/a

Data analysis

n/a

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The authors declare that the data supporting the findings of this study are available within the paper [and its supplementary information].

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study is a multidisciplinary research effort involving analysis and interpretation of prehistoric rock art in southern Sulawesi, Indonesia, as well as its direct-dating using Uranium-series analysis of overlying calcium carbonate materials (coralloid speleothem).
Research sample	The research sample comprises a total of four coralloid speleothem deposits directly associated with three animal motifs from a single rock art panel at a limestone cave (Leang Bulu' Sipong 4) in the Maros-Pangkep karst region of southwestern Sulawesi.
Sampling strategy	Opportunistic - when coralloid speleothems or other calcite deposits deemed to be of sufficient quality for Uranium-series dating were found in direct association with rock art motifs relevant to the study topic we collected them as dating samples.
Data collection	Maxime Aubert collected the coralloid speleothems following procedures outlined in detail in the Methods section of the paper.
Timing and spatial scale	The cave art site under study was discovered and recorded by our team in December 2017. The coralloid speleothem samples were collected by Maxime Aubert in February 2018.
Data exclusions	n/a
Reproducibility	n/a
Randomization	n/a
Blinding	n/a
Did the study involve field work?	<input checked="" type="checkbox"/> Yes <input type="checkbox"/> No

Field work, collection and transport

Field conditions	Leang Bulu' Sipong 4 is located in a near-coastal lowland tower karst region (Maros-Pangkep). The field area is near-equatorial and thus hot and humid for most of the year. Our field research was conducted during the driest part of the year (June to September), with the exception of the February 2018 sample collection trip which was conducted during the monsoonal wet season.
Location	The general location of the site is indicated in Figure 1 in the paper but specific co-ordinates are not provided to protect it from unauthorised visits, vandalism, etc.
Access and import/export	Samples were exported to Australia for dating under the material transfer agreement of the Memorandum of Understanding between the Indonesian scientific counterpart (ARKENAS) and Griffith University, and also following collaborative arrangements with other local stakeholders.
Disturbance	We removed small calcium carbonate deposits that had formed over the top of, and thus partially obscured, rock art motifs

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Frontal cortex neuron types categorically encode single decision variables

<https://doi.org/10.1038/s41586-019-1816-9>

Junya Hirokawa^{1,2,5}, Alexander Vaughan^{1,5}, Paul Masset^{1,3,4}, Torben Ott¹ & Adam Kepecs^{1*}

Received: 8 May 2017

Accepted: 15 October 2019

Published online: 4 December 2019

Individual neurons in many cortical regions have been found to encode specific, identifiable features of the environment or body that pertain to the function of the region^{1–3}. However, in frontal cortex, which is involved in cognition, neural responses display baffling complexity, carrying seemingly disordered mixtures of sensory, motor and other task-related variables^{4–13}. This complexity has led to the suggestion that representations in individual frontal neurons are randomly mixed and can only be understood at the neural population level^{14,15}. Here we show that neural activity in rat orbitofrontal cortex (OFC) is instead highly structured: single neuron activity co-varies with individual variables in computational models that explain choice behaviour. To characterize neural responses across a large behavioural space, we trained rats on a behavioural task that combines perceptual and value-guided decisions. An unbiased, model-free clustering analysis identified distinct groups of OFC neurons, each with a particular response profile in task-variable space. Applying a simple model of choice behaviour to these categorical response profiles revealed that each profile quantitatively corresponds to a specific decision variable, such as decision confidence. Additionally, we demonstrate that a connectivity-defined cell type, orbitofrontal neurons projecting to the striatum, carries a selective and temporally sustained representation of a single decision variable: integrated value. We propose that neurons in frontal cortex, as in other cortical regions, form a sparse and overcomplete representation of features relevant to the region's function, and that they distribute this information selectively to downstream regions to support behaviour.

The brain represents the external world in patterns of neural activity that guide adaptive behaviour. In many regions, individual cortical neurons respond to features, such as visual edges¹, objects² or spatial locations³, that reflect regional function. When examining frontal areas engaged in decision making, however, one is struck most of all by the complexity and diversity of their neuronal responses^{4–14,16–18}. The difficulty in identifying structure in frontal cortical representations probably reflects the fact that cognitive variables are more challenging to define than simpler features such as visual edges. It is further challenging to design behavioural tasks that engage the specific cognitive functions of frontal cortical neurons to sufficiently probe the relevant feature space. As a result, it is unclear whether frontal cortex representations are comprehensible in single neurons or instead require neural population analysis (for example, ‘random mixed selectivity’^{14,15}, but see ref.¹⁹).

Behavioural task engaging orbitofrontal cortex

We trained Long Evans rats on a complex reward-biased psychometric olfactory discrimination task (Fig. 1a) that requires integration

of decision confidence and reward value. This task separately varies perceptual uncertainty across trials and reward expectations across blocks. Choice accuracy varied systematically with stimulus difficulty; changing reward size induced rapid and sustained changes in choice biases and reaction times²⁰ (Fig. 1b, c and Extended Data Fig. 1a–e; see Methods), confirming that animals' strategy maximized total reward (Extended Data Fig. 1f, g; see Methods). A model encoding key decision variables accounts for overall choice strategy (Fig. 1d, e) and trial-by-trial biases arising from previous outcomes (Extended Data Fig. 1h–l).

We recorded 485 neurons from lateral orbitofrontal cortex (OFC) in three rats (Extended Data Fig. 2) and analysed the post-choice epoch when rats await an uncertain reward (‘reward anticipation’, Fig. 1a). We identified OFC neurons that encode canonical variables in our model, including decision confidence, with activity proportional to the evidence supporting the choice, but not influenced by expected reward size^{8,21} (Fig. 2a, b). We also observed representations of other canonical decision variables, including anticipated reward size (assuming a correct choice) and integrated value (that is, probability of reward for a given choice multiplied by reward size; Extended Data Fig. 3a).

¹Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. ²Graduate School of Brain Science, Doshisha University, Kyotanabe, Kyoto, Japan. ³Watson School of Biological Sciences, Cold Spring Harbor, NY, USA. ⁴Present address: Department of Molecular and Cellular Biology, Centre for Brain Science, Harvard University, Cambridge, MA, USA. ⁵These authors contributed equally: Junya Hirokawa, Alexander Vaughan. *e-mail: kepecs@cshl.edu

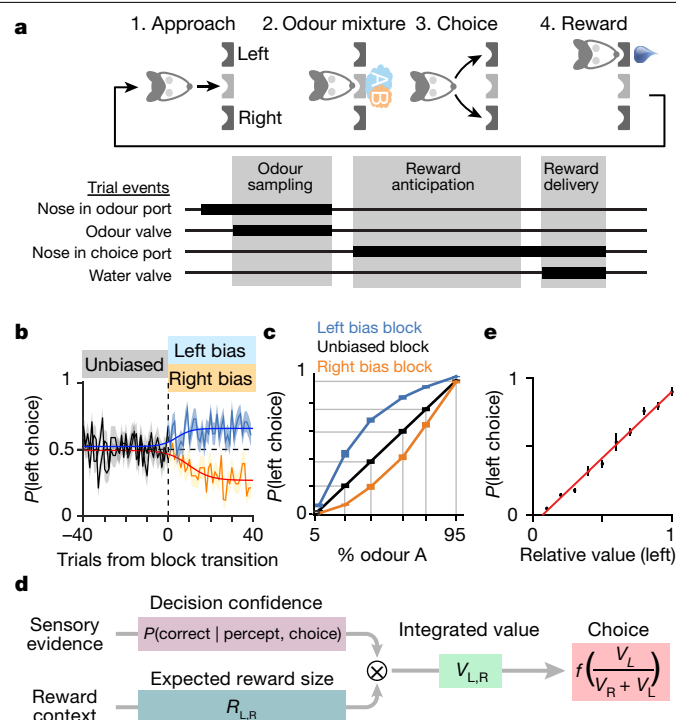


Fig. 1 | Reward-biased psychometric odour discrimination task to probe decision-variable integration. **a**, Task design, single trial. We primarily analysed the 1 s reward anticipation window. **b**, Behavioural performance is modified by block-wise changes in reward size. Choice biases emerged rapidly across blocks (5.45 ± 0.48 trials to bias shift, mean \pm s.e.m., $n = 67$ sessions from 3 rats). **c**, Average psychometric functions in unbiased, left- and right-bias blocks reveal systematic bias (30 sessions from one rat (C068)). **d**, Decision-variable model of choice behaviour in this task integrates variables from reward context (choice, size of potential reward, presence of reward) and sensory evidence to generate internal variables representing expected reward value and decision confidence. These combine to estimate integrated value, which determines choice probability. **e**, The relative value derived from the model explains choice behaviour (mean \pm s.e.m., $n = 67$ sessions from 3 rats).

Clustered single neuron representations

We next considered whether these OFC neurons were representative or outliers from a diverse population (Fig. 2b–g). We generated response profiles for each neuron across 42 task contingencies combining odour stimulus, reward size, behavioural choice, and previous trial outcome (Fig. 2b and Extended Data Fig. 3b–d). We developed two improved statistical tests to examine deviations from random mixed selectivity (ePAIRS and eRP) and showed that OFC neurons do not encode random mixtures of available information^{22,23} (Extended Data Fig. 4a–i; see Methods). Instead, similarly tuned neurons are more common than expected (Fig. 2d, e and Extended Data Fig. 4e–m), and neuronal activity does not uniformly fill the space of available representations (Extended Data Fig. 3e, f).

We used a model-free approach to identify clusters of OFC neurons with similar response profiles. To alleviate the challenges of a high-dimensional response space, we used nonlinear spectral clustering with bootstrap validation to identify clusters as strongly connected subcomponents of a neighbourhood graph (Fig. 2c; Methods). This analysis identified ~9 robust clusters (Fig. 2f), each encoding one ‘categorical’ representation (Fig. 2g).

As controls, we confirmed that clustering did not arise from segregation of neurons from different rats (Extended Data Fig. 5a, b), spatial patterning in the OFC (Supplementary Note 1 and Extended Data Fig. 5c–h), or data pre-processing (Extended Data Fig. 5i, j; see Methods). Additionally, our approach was biased against finding such clusters in several ways (see Supplementary Discussion).

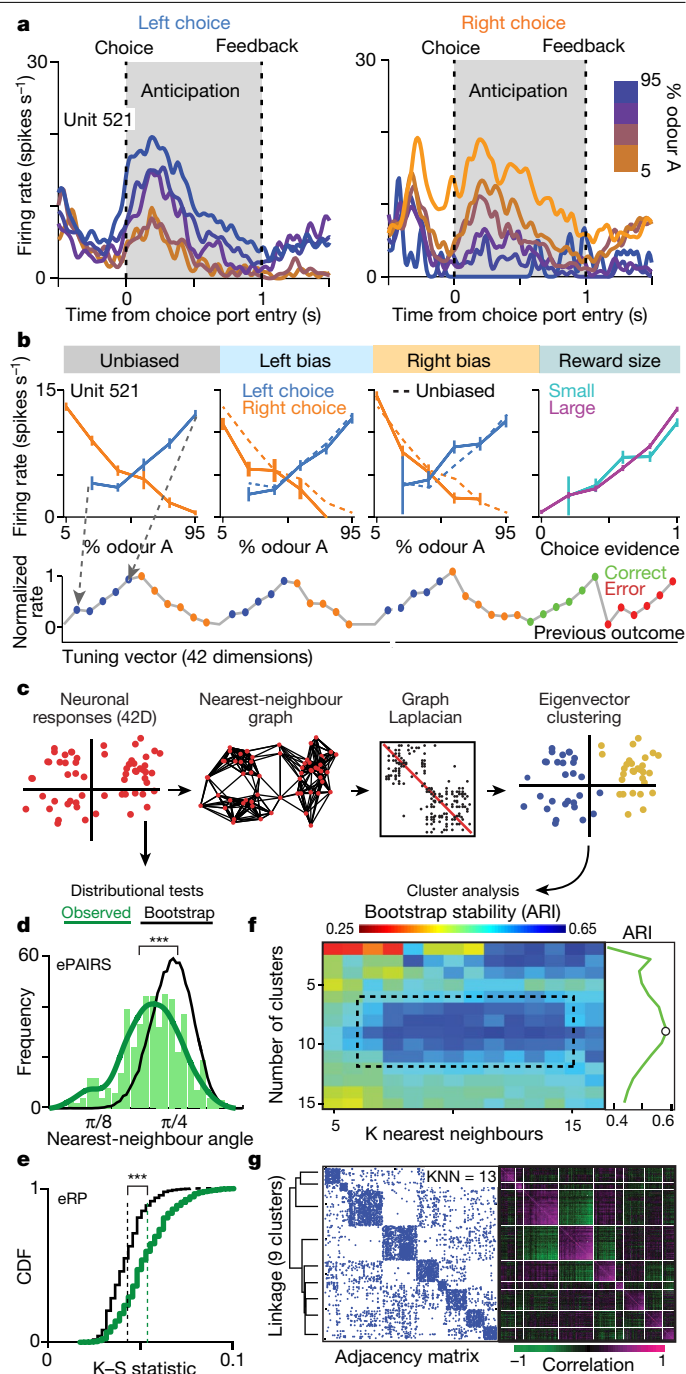


Fig. 2 | OFC neurons form nine discrete clusters. **a**, **b**, Response during reward anticipation for one neuron (unit 521). Activity in the unbiased block reflected evidence supporting animal’s choice (**a**), and matched expected response profile of representation of decision confidence (varying with choice evidence, but not reward size) (**b**). **c**, Analysis pipeline for cluster analysis. **d**, **e**, Distribution of nearest-neighbour angles (**d**) and projection angles (**e**) reveal non-random distribution in OFC population with angle magnitudes suggesting clustering ($***P < 0.001$ for both tests; Methods, Extended Data Fig. 4). **f**, Spectral clustering reveals nine robust groups of neurons, with high cluster stability across a range of hyperparameters (dashed box; adjusted rand index, ARI; see Methods for details of cross-validation). **g**, Sorted adjacency and correlation matrices reveal strong within-cluster similarity and between-cluster antagonism. CDF, cumulative distribution function; K-S, Kolmogorov–Smirnov; KNN, K nearest neighbours.

To confirm reproducibility, we repeated our complete analysis on an independent cohort of four rats, and observed categorical selectivity in OFC neurons with only minor differences (639 neurons; Extended

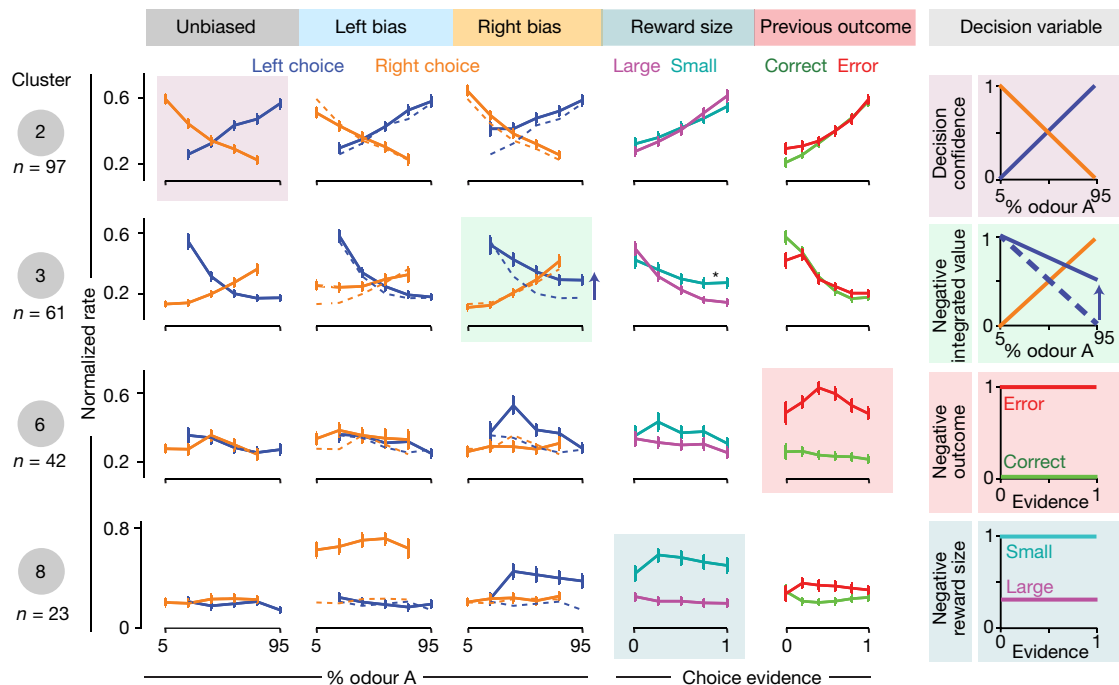


Fig. 3 | OFC neuron clusters represent putative decision variables. Average response profiles for 4 out of 9 clusters (see Extended Data Fig. 7e for others). Each row represents a cluster, with average normalized firing rate plotted across behavioural conditions. The first three conditions represent combinations of stimulus and choice in different reward blocks (unbiased, left, right bias). Reward size column represents tuning as a function of choice evidence and reward size blocks (left and right bias). Previous outcome column

represents tuning as a function of choice evidence and previous trial outcome (correct rewarded, error not rewarded). Tuning profiles suggest specific decision variables represented. For each cluster, we highlight the expected shape of response profile that indicates which putative decision variable is represented, including potential reward-size biases that differ from the non-biased condition (dashed line). Error bars show s.e.m. * $P < 0.01$, t -test.

Data Fig. 4m, Extended Data Fig. 6 and Supplementary Note 2 'Cohort analysis').

Decision variable coding

Although clusters were identified without model input, the average response profile of each cluster resembles a putative decision variable in the behavioural model (Fig. 3), including decision confidence (confidence⁽⁺⁾, cluster 2), integrated value (value⁽⁻⁾, cluster 3), previous trial outcomes (previous outcome⁽⁻⁾, cluster 6), reward size (cluster 8), and others (Extended Data Fig. 7).

We confirmed this correspondence by analysing three clusters—confidence⁽⁺⁾, confidence⁽⁻⁾ and integrated value⁽⁻⁾—for which we had quantitative predictions. Confidence⁽⁺⁾ neuron (cluster 2) responses matched key expectations of statistical decision confidence^{8,21,24} (Fig. 4a and Extended Data Fig. 8a). Neural activity increased with stimulus contrast for correct choices but decreased for errors (Fig. 4a, bottom left), and firing rates predicted choice accuracy regardless of stimulus identity (Fig. 4a, bottom right). The confidence⁽⁻⁾ cluster showed the same characteristics, sign-reversed (Extended Data Fig. 8b). Integrated value⁽⁻⁾ neurons (cluster 3, Fig. 4b) also matched predictions, representing confidence and reward size in population averages (Fig. 4b, bottom left and Extended Data Fig. 8c) and correlates strongly with model estimates in individual neurons (Extended Data Fig. 8c (panels ii and vi)). Critically, firing rate tracked behavioural accuracy below 50%, reflecting outcome probability for decisions made under reward bias (Fig. 4b, bottom right).

We next considered how well the OFC population reflects decision variables without specifying a particular model. We generated a canonical regression model of elementary task variables (for example, odour stimulus, choice side, expected reward) that combine to

produce decision variables (for example, confidence, integrated value; Fig. 4c, Methods and Extended Data Fig. 8e). Individual OFC neurons and cluster-averaged responses were well fit by this elementary model (clusters, $P < 1 \times 10^{-64}$; neurons, $P < 1 \times 10^{-10}$; t -test, Fig. 4d), revealing that each neuron can be represented in a space spanned by task-relevant variables.

Nevertheless, responses in this linear model may still represent arbitrary mixtures of task variables, rather than identifiable decision variables. As OFC neurons tend to represent coherent variables like integrated value⁽⁺⁾ (combining confidence⁽⁺⁾ and reward size⁽⁺⁾; Extended Data Fig. 8d) rather than incoherent combinations (for example, confidence⁽⁺⁾ and reward size⁽⁻⁾), we tested whether representations of canonical decision variables were enriched in OFC. To do so, we compared regressions of OFC responses using the canonical model against a library of models containing the same variables randomly mixed by basis set rotation (Fig. 4e; see Methods). Although both models can represent any combination of variables, only the canonical model represents model-based decision variables sparsely. As expected, LASSO regression of cluster-averaged response profiles using the canonical model had higher sparsity (smaller L1 penalty, $P < 0.00042$) without penalizing fit ($P > 0.05$; Fig. 4f–g and Extended Data Fig. 8f). We observed similar results for individual neurons (Fig. 4h and Extended Data Fig. 8g). Thus, OFC representations identified by model-free clustering correspond to the canonical elements of a decision-variable model.

We observed similarly robust clustering in other epochs of the behavioural task, albeit with fewer clusters arising from the smaller number of behavioural variables in the stimulus and feedback epochs (Extended Data Fig. 7). Additionally, we observed structured transitions in decision-variable coding across epochs: neurons within a cluster were likely to co-cluster also in other epochs (Fig. 5a, Extended Data

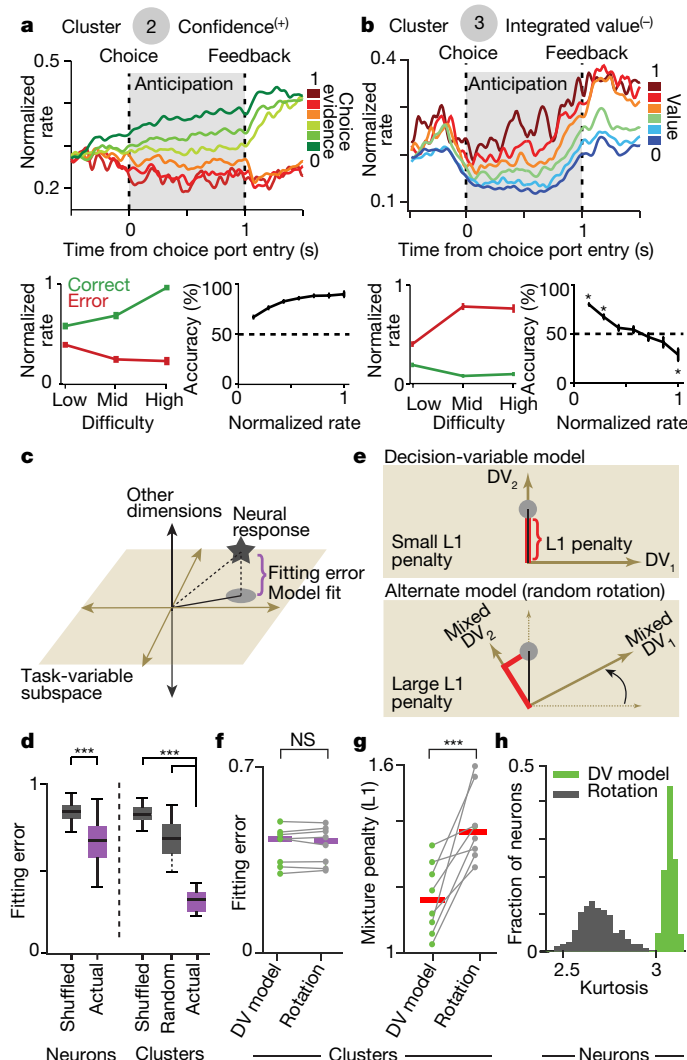


Fig. 4 | OFC clusters quantitatively encode the full decision-variable model. **a**, Response profile of neurons in cluster 2 correspond to decision variable confidence⁽⁺⁾. Top panel shows peri-stimulus time histogram of normalized firing rate, grouped by evidence supporting choice. Bottom left panel shows firing rate as a function of stimulus difficulty and choice. Bottom right panel shows choice accuracy as a function of firing rate. **b**, Same convention as panel **a** for cluster 3, corresponding to decision variable integrated value⁽⁻⁾. **c**, Cartoon of decision-variable model regression fit. **d**, Mean response profiles of single neurons and averaged clusters were fit: error was lower for actual neurons than for trial-shuffled control, and dramatically lower for fits of cluster averages than for single neurons, clusters of trial-shuffled neurons, or random clusters. Error distributions were compared under cross-validation by bootstrap test. Error bars show s.e.m., with horizontal links showing significant differences. *** $P < 0.001$. **e**, Least absolute shrinkage and selection operator (LASSO) regression analysis of individual clusters. We used two sets of decision-variable models to fit average cluster response profile: canonical model using elementary task variable such as stimulus and choice side, and alternate models in which these elementary task variables are randomly mixed (for example, choice side + reward block type – stimulus evidence). **f**, **g**, Fits of canonical decision-variable model (DV model) to clusters required smaller mixture penalty (L1, *** $P < 0.001$, sign rank test) than rotated variables (that is, mixtures of decision variables), with similar fitting error. NS, not significant; horizontal lines show tested comparisons across groups. **h**, Individual neurons are also better fit to canonical decision variables than their rotated versions: distribution of regression coefficients derived from fitting individual neuronal responses has higher kurtosis when fit with the canonical decision-variable model than with rotated models ($P < 0.001$; bootstrap test).

Fig. 7 and Supplementary Note 3). An analogous clustering procedure focused on temporal features of neural activity revealed robust structure in neuronal dynamics, with most neurons sparsely activated in a few task epochs and clustering into ~8 temporal profiles (Fig. 5b and Extended Data Fig. 9a–e).

Cell-type-specific coding

Seeking an anatomical substrate for these response profiles, we examined OFC neurons projecting to striatum (OFC–STR), a pathway important for reversal learning and choice value updating^{25,26}. We used retrograde viruses to target ChR2 to OFC–STR neurons and identified 24 photo-tagged neurons based on signatures of direct light activation (Fig. 5c–e and Extended Data Fig. 10). Notably, their behavioural tuning and time course matched specific representations identified by model-free clustering (Fig. 5b). OFC–STR neurons significantly encoded trial outcome (23 out of 24, 5 positive, 18 negative), with activity sustained beyond the feedback period until the next trial. Negatively tuned OFC–STR neurons reduced their activity during the anticipation period (Fig. 5f, g), encoding negative integrated value. They increased their firing after negative outcomes during the feedback epoch, and sustained firing throughout the self-paced inter-trial interval, often for many seconds. This neuronal profile matched the dynamics of one temporal cluster, with the same transition from negative integrated value to sustained negative outcome coding (Fig. 5h, $n = 96$ neurons, cluster B in Fig. 5b). These results, including a similar pattern for positively tuned OFC–STR neurons (Extended Data Fig. 9f, g), suggest that the temporally structured and decision-variable-specific representations in OFC are supported, at least in part, by cell-type-specific circuit organization.

Representational logic of frontal cortex

We used a behavioural task that combines perceptual and value-guided decision making to demonstrate that OFC representations are highly structured: encoding a small set of categorical representations that correspond to coherent decision variables in specific task epochs. The functionally homogeneous encoding of decision variables by OFC–STR neurons further suggests that OFC response diversity is partly due to cell-type-specific organization. Sustained firing of OFC–STR neurons across trials encoded value, potentially a neural correlate of a temporal credit assignment mechanism. Alongside neuron-type-specific recordings in cortical and subcortical areas during behaviour^{27–29}, this result exposes an intimate connection between the functional and anatomical organization of cortex supporting computation.

The internal decision variables guiding behaviour (for example, value) can only be indirectly determined, like identifying an object from its shadow, by conditioning neural activity on measurable task variables (for example, stimulus, choice). Consequently, categorical representations of internal variables like decision confidence can appear mixed when examined as a function of external variables (Fig. 3). Our unbiased, model-free approach revealed that many hypothesized decision variables have a privileged but possibly task-dependent representation, constraining models of OFC function and clarifying its representational logic.

This structured representation of decision variables has strong analogies to the framework of sparse and overcomplete representations providing efficient sensory encoding³⁰. OFC shares many similarities with these regions: sparse activation (Extended Data Fig. 3e, f), redundant encoding (Figs. 2 and 3) and representational sparsity (Fig. 4). We propose that this architecture is a fundamental feature of frontal cortex, with distinct cell types in OFC specializing in different computational functions to support adaptive behaviour.

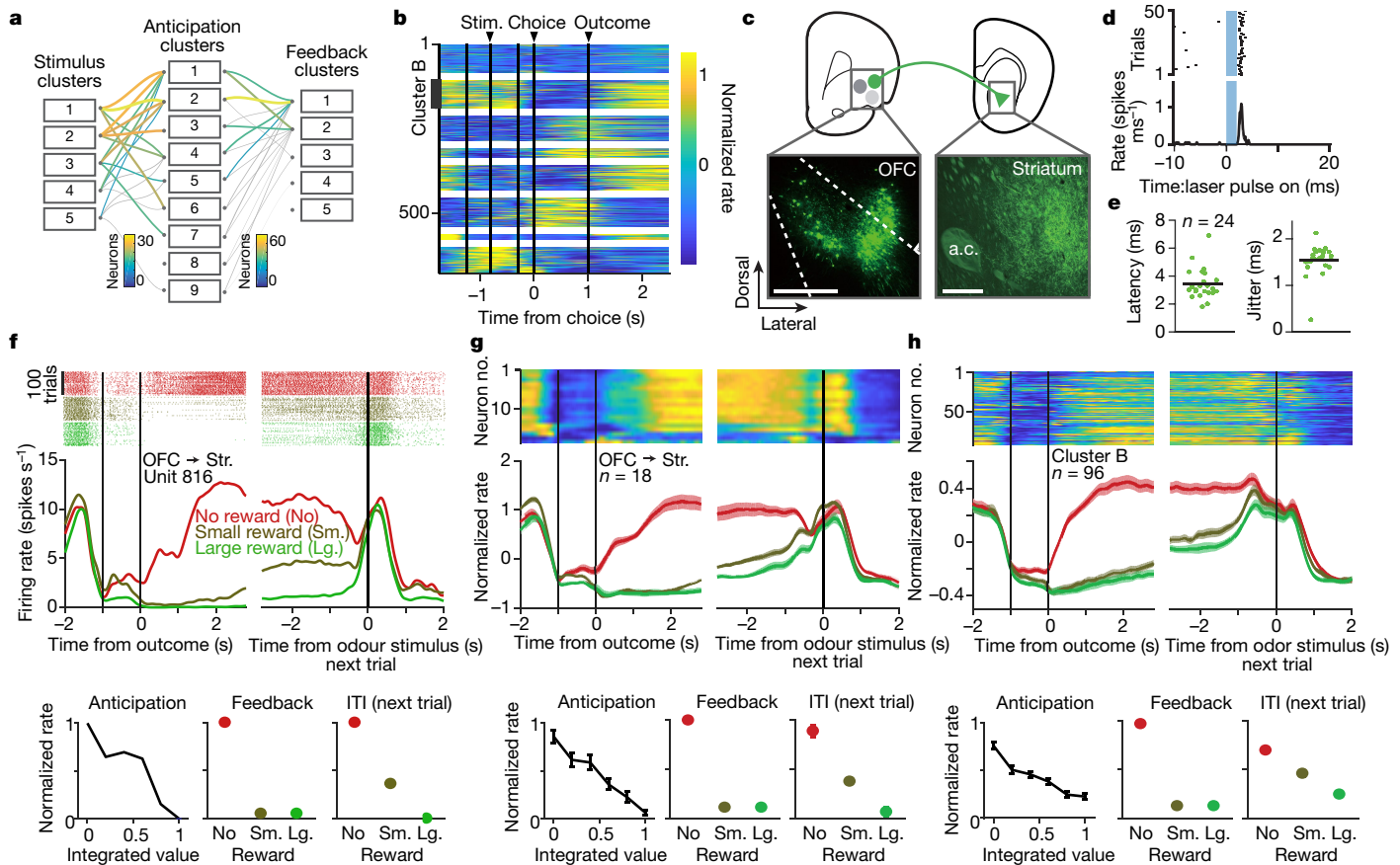


Fig. 5 | OFC-to-striatum projection neurons encode integrated value across task epochs. **a**, Cluster transitions across epochs: each box represents a neuron cluster in one epoch; lines between them represent neurons that belong to the same cluster in two epochs. Transitions in cluster identities across trial epochs are highly structured, with two neurons in a given cluster likely to remain in the same clusters over time. **b**, Dynamics of response time course for eight temporal clusters (unrewarded trials, see Extended Data Fig. 9d). **c**, Retrograde targeting of channelrhodopsin-2 (ChR2) to striatum-projecting neurons in lateral OFC for optogenetic identification (white dashed lines indicate approximate border for lateral OFC), scale 500 μm . **d**, Laser pulse (blue band) aligned responses of an optogenetically identified neuron. **e**, Latency and timing jitter of spikes from laser pulse onset for identified OFC-STN neurons was short and low (horizontal bars show averages). **f**, Spike raster and average activity in single striatum-projecting neuron aligned to outcome

(reward or error sound, left) or to stimulus delivery in the next trial (right). Lower panels show analysis of firing rates in three different epochs: anticipation, for which firing rate varies with integrated value (left); feedback, for which firing rate reflects trial outcome with negative valence (middle); inter-trial interval (ITI), for which firing rates reflect graded outcomes with negative valence (right; NR, no reward; SR, small reward; LR, large reward). **g**, Average activity in error trials for optogenetically identified OFC-STN projection neurons (negative outcome selective, 18 of 24, area under the curve (AUC) < 0 with $P < 0.01$ permutation test) and their average peri-stimulus time histogram (PSTH) grouped by error, correct small and large reward trials. **h**, Average dynamics and tuning of neurons in temporal cluster B (panel **b**) match those of optogenetically identified striatum-projecting neurons in **f, g**. To avoid double counting, identified OFC-STN projection neurons were excluded from cluster B. Error bars represent mean \pm s.e.m. in **g** and **h**.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1816-9>.

- Hubel, D. H. & Wiesel, T. N. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* **148**, 574–591 (1959).
- Bruce, C., Desimone, R. & Gross, C. G. Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *J. Neurophysiol.* **46**, 369–384 (1981).
- O'Keefe, J. & Dostrovsky, J. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* **34**, 171–175 (1971).
- Abe, H. & Lee, D. Distributed coding of actual and hypothetical outcomes in the orbital and dorsolateral prefrontal cortex. *Neuron* **70**, 731–741 (2011).
- Feierstein, C. E., Quirk, M. C., Uchida, N., Sosulski, D. L. & Mainen, Z. F. Representation of spatial goals in rat orbitofrontal cortex. *Neuron* **51**, 495–507 (2006).
- Roesch, M. R., Taylor, A. R. & Schoenbaum, G. Encoding of time-discounted rewards in orbitofrontal cortex is independent of value representation. *Neuron* **51**, 509–520 (2006).

- Kennerley, S. W. & Wallis, J. D. Evaluating choices by single neurons in the frontal lobe: outcome value encoded across multiple decision variables. *Eur. J. Neurosci.* **29**, 2061–2073 (2009).
- Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227–231 (2008).
- Machens, C. K., Romo, R. & Brody, C. D. Functional, but not anatomical, separation of “what” and “when” in prefrontal cortex. *J. Neurosci.* **30**, 350–360 (2010).
- McGinty, V. B., Rangel, A. & Newsome, W. T. Orbitofrontal cortex value signals depend on fixation location during free viewing. *Neuron* **90**, 1299–1311 (2016).
- Morrison, S. E. & Salzman, C. D. The convergence of information about rewarding and aversive stimuli in single neurons. *J. Neurosci.* **29**, 11471–11483 (2009).
- Padoa-Schioppa, C. Neuronal origins of choice variability in economic decisions. *Neuron* **80**, 1322–1336 (2013).
- Padoa-Schioppa, C. & Assad, J. A. Neurons in the orbitofrontal cortex encode economic value. *Nature* **441**, 223–226 (2006).
- Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
- Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
- Stalnaker, T. A., Cooch, N. K. & Schoenbaum, G. What the orbitofrontal cortex does not do. *Nat. Neurosci.* **18**, 620–627 (2015).
- Steiner, A. P. & Redish, A. D. Behavioral and neurophysiological correlates of regret in rat decision-making on a neuroeconomic task. *Nat. Neurosci.* **17**, 995–1002 (2014).

18. Sul, J. H., Kim, H., Huh, N., Lee, D. & Jung, M. W. Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. *Neuron* **66**, 449–460 (2010).
19. Xie, J. & Padoa-Schioppa, C. Neuronal remapping and circuit persistence in economic decisions. *Nat. Neurosci.* **19**, 855–861 (2016).
20. Zariwala, H. A., Kepecs, A., Uchida, N., Hirokawa, J. & Mainen, Z. F. The limits of deliberation in a perceptual decision task. *Neuron* **78**, 339–351 (2013).
21. Hangya, B., Sanders, J. I. & Kepecs, A. A Mathematical framework for statistical decision confidence. *Neural Comput.* **28**, 1840–1858 (2016).
22. Raposo, D., Kaufman, M. T. & Churchland, A. K. A category-free neural population supports evolving demands during decision-making. *Nat. Neurosci.* **17**, 1784–1792 (2014).
23. Cuesta-Albertos, J. A., Cuevas, A. & Fraiman, R. On projection-based tests for directional and compositional data. *Stat. Comput.* **19**, 367–380 (2009).
24. Sanders, J. I., Hangya, B. & Kepecs, A. Signatures of a statistical computation in the human sense of confidence. *Neuron* **90**, 499–506 (2016).
25. Morris, L. S. et al. Fronto-striatal organization: Defining functional and microstructural substrates of behavioural flexibility. *Cortex* **74**, 118–133 (2016).
26. Burguière, E., Monteiro, P., Feng, G. & Graybiel, A. M. Optogenetic stimulation of lateral orbitofronto-striatal pathway suppresses compulsive behaviors. *Science* **340**, 1243–1246 (2013).
27. Economo, M. N. et al. Distinct descending motor cortex pathways and their roles in movement. *Nature* **563**, 79–84 (2018).
28. Kvitsiani, D. et al. Distinct behavioural and network correlates of two interneuron types in prefrontal cortex. *Nature* **498**, 363–366 (2013).
29. Nambodiri, V. M. et al. Single-cell activity tracking reveals that orbitofrontal neurons acquire and maintain a long-term memory to guide behavioral adaptation. *Nature Neurosci.* **22**, 1110–1121 (2019).
30. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

Rats

Nine male Long Evans rats (~300 g) were used for the study. Seven rats were purchased from Taconic Biosciences (NY, USA) and two rats were obtained from Shimizu Laboratory Supplies (Kyoto, Japan). The rats were pair-housed and maintained on a reverse 12 h light/dark cycle and tested during their dark period. Food was available ad libitum, and rats had scheduled access to water with daily body weight monitoring to ensure that body mass remained above 85% of initial weight. Rats received water during behavioural sessions, and ad libitum in the following 30 min in their home cage. All procedures involving animals were carried out in accordance with National Institutes of Health standards and were approved by the Cold Spring Harbour Laboratory Institutional Animal Care and Use Committee and by the Animal Research Committee of Doshisha University.

Behavioural apparatus

The behavioural setup consisted of three ports equipped with infrared photodiodes; interruption of the infrared beam signalled port entry⁸. Odours were mixed with pure air to produce a 1:10 dilution at a flow rate of 1 l min⁻¹ using a custom-built olfactometer (Island Motion; AALBORG). Delivery of odours and water reinforcement was controlled using Pulse Pal³¹ and Bpod (J. Sanders and A.K.) with custom software written in MATLAB. Water rewards were delivered from gravity-fed reservoirs regulated by solenoid valves 1 s after the subject entered the choice port.

Reward biased odour discrimination task

Rats were trained and tested on a reward-biased two-alternative forced choice (2AFC) odour discrimination task as follows. Rats self-initiated each trial with a nose-poke into the central port to receive the odour stimulus. After a variable delay (uniform distribution of 0.2–0.6 s), a binary mixture of two pure odorants (*S*(+)-2-octanol and *R*(-)-2-octanol) was delivered at one of six concentration ratios ranging from 5% to 95% in pseudorandom order within a session. The specific odour ratios delivered were varied during training to produce graded accuracy levels from chance to near perfect performance.

After a variable odour sampling time, rats responded by withdrawing from the central port, thus terminating odour delivery, and moved to the left or right choice ports. Choices were rewarded according to the dominant component of the mixture, that is, at the left port for mixtures $A/B > 50/50$ and at the right port for $A/B < 50/50$. Reward amount was set to 0.025 ml in the control blocks and was reduced to 0.3–0.5 of the original amount for either right or left in left-biased reward block or right-biased reward block, respectively. The specific reward bias was determined to encourage graded responses to the reduced reward-size side and avoid excessive bias towards the larger reward side. Error choices resulted in water omission and were signalled by white noise (Fig. 1a, b; except for cohort 2 rats, see Supplementary Note 2 ‘Cohort analysis’). Two additional rats (cohort 3) were tested in a version of the task without reward size manipulation. In each cohort, randomization was not required as all rats were treated similarly.

In order to maximally utilize the psychometric response space, however, we intentionally modified the set of odour concentrations used in each session to generate a linearized set of choice probabilities (approximately [0, 0.2, 0.4, 0.6, 0.8, 1.0]) in the control block (that is, when reward sizes are equal; Fig. 1c). This linearized odour representation is reported using the axis legend ‘% odour A’ in relevant figures.

Training

Rats were trained over the course of 4–6 weeks, with progressive introduction of each aspect of the task: imperative trials (not conditional on stimulus) followed by choice trials (conditional on the stimulus). Both trial difficulty and the delay before reward delivery were gradually

increased until behaviour was stable; subsequently, reward bias was gradually introduced by expanding reward amount differences from baseline.

Surgery

Anaesthesia was induced with inhalation of 2.5% isoflurane and retained with intraperitoneal injections of ketamine (50 mg kg⁻¹) and medetomidine (0.4 mg kg⁻¹) at the onset of the surgery and supplemented as necessary based on the hind leg reflex. Body temperature was maintained using a heating pad (HoMedics). Rats were stereotactically implanted with custom-made microdrives in the left orbitofrontal cortex (targeted 1.5 mm above OFC (AP +3.7, ML ±3.2, DV +3.0; AP, anterior–posterior; ML, medial–lateral; DV, dorsal–ventral). Following surgery, rats were administered ketoprofen (Fort Dodge Animal Health) as an analgesic (5 mg kg⁻¹). Rats were monitored during their recovery from surgery at least 7 days before recordings began.

Electrophysiological recordings

Custom-built light-weight microdrives were constructed for deep brain recording and optogenetic stimulation with an optic fibre and 8 tetrodes. 3D printed microdrive bodies housed moveable shuttles that converted screw rotation into vertical motion advancing the shuttle and with it the attached tetrode and/or optical fibre. Individual tetrodes consisted of four twisted polyimide-coated nichrome wires (Precision Fine Tetrode Wire, Sandvik; single wire diameter 12 µm, gold plated to 0.25–0.5 MΩ impedance). Extracellular recordings were acquired on a DigitalLynx data acquisition system (Neuralynx) with a sampling rate of 31.25 kHz. Tetrodes were advanced daily (approximately 90 µm) after recording sessions so as to sample an independent population of cells across sessions and all the recorded neurons were analysed. For cohort 3 (2 rats, HA56 and HA58), the OpenEphys system was used for recording with a sampling rate of 30 kHz.

Histology

Rats were anaesthetized (sodium pentobarbital; overdose) and then transcardially perfused with saline and 4% paraformaldehyde. The brains were removed and 100 µm serial coronal sections were prepared with a vibratome. Recording sites were marked by coating electrodes with fluorescent dye (Vybrant DiI, Invitrogen) or by electrolytic lesions.

Viral injection

To target striatum-projecting OFC neurons we used a retrograde labelling approach. For rats CO91, S11 and HA56, we used adeno-associated virus (AAV) 2/9 serotype (5E12 pp ml⁻¹ (physical particle ml⁻¹) UNC Vector Core Facility) carrying EF1a-DIO-ChR2-EYFP³² or hSyn-DIO-{mCAR}_{off}{ChR2}_{on} (ref. ³³) and CAV2-Cre 4.1E12 pp ml⁻¹ (ref. ³⁴) (Montpellier vectorology platform), injected on the same day into 4-week-old rats in OFC and striatum, respectively. For rat HA58 AAV retro³⁵, AAV-Syn-ChR2(H134R)-GFP (AddGene) was injected in the striatum. Rats were anaesthetized using 2.5% isoflurane and their eyes were protected with ophthalmic lubricant (Puralube Vet Ointment, Dechra Pharmaceuticals). Rats were placed in a stereotaxic apparatus (David Kopf Instruments, or Narishige) and the skull was levelled along both the antero-posterior and medio-lateral axis to allow the target coordinates: OFC (AP, +3.7 mm; ML, 3.2 mm; DV, 3.0 and 2.7 mm) and striatum (AP, +1.2 mm; ML, 2.3 mm; DV, 6.3 and 6.0 mm). At each dorso-ventral level, 150 nl virus was injected slowly for 4 min via a glass pipette pulled (P-97 Flaming/Brown Micropipette Puller, Sutter Instruments) from borosilicate capillaries (5 µl; tip diameter 20 µm). Injections were carried out by delivering brief pulses of pressure using Picospritzer II (Parker) or a micropump (UltraMicroPump III, WPI). After injections, the pipette was slowly pulled out after a 5 min waiting time. Rats were housed in their home cage and training commenced once they reached 14 weeks of age.

Optical stimulation

Optical stimulation was performed as previously described²⁸. In brief, a multimode optical fibre (55 µm diameter NA = 0.22, Polymicro Technologies) was coupled via a modified LC–LC type connector to a multimode fibre (126 µm diameter NA = 0.27, CablesPlus), which collected light from a blue laser (473 nm; 50 mW; Ultralasers). Maximal power at the tip of the fibre ranged from 6 to 15 mW of total output. The light stimulation protocol (15–30 min) for optogenetic tagging was performed at the end of each recording session consisting of varying frequencies (10, 20 and 40 Hz with 1 ms or 3 ms light pulses) and intensities (0.1–10 mW) to enable reliable identification of directly light-activated neurons (Extended Data Fig. 10).

Behavioural data analysis

Each trial was defined by the stimulus presented, the animal's choice, and the reward associated with each choice. We calculated the odour sampling duration (OSD) as the difference between odour valve actuation and the odour port exit, with 100 ms subtracted to account for the delay from valve opening to odour reaching the nose (Fig. 1a and Extended Data Fig. 1e, l).

Daily behavioural sessions consisted of 821 ± 11 (mean \pm s.e.m.) trials for cohort 1 (67 sessions), 893 ± 33 for cohort 2 (42 sessions) and 683 ± 19 for cohort 3 (photo-tagged rats, 47 sessions). For each session, we calculated behavioural accuracy as the fraction of correct choices, excluding trials in which OSD was less than 100 ms (0.47% of all trials) and trials in which no choice was made before the trial was terminated (0.82% of trials). Error bars are mean \pm s.e.m. (n across rats) or mean \pm s.d. (n across sessions).

The effect of odour contrast on accuracy or OSD was tested using one-way ANOVA with pairwise comparisons between different mixture contrast ratios at a significance level of $P < 0.0125$ (that is, adjusted for multiple comparisons).

Behavioural choice model

We developed a computational model for our behavioural task to describe the integration of sensory and reward information by the animal. In doing so, we aimed to predict choice behaviour for both biased and unbiased blocks, under four assumptions: (1) we assumed that animals rapidly learned and maintained a constant reward size estimate for each choice option within a given reward block. This is justified by the observation that animals adapted to new bias blocks within ~5 trials (Fig. 1b). (2) We assumed that animals engaged in our sensory discrimination task exploited the odour mixture to generate a choice (estimating the dominant mixture) but also to generate an internal estimate of outcome likelihood (correct or error) for each choice option (left or right)⁸. (3) We assumed that relative reward estimates, R , and outcome likelihood, $L = P(\text{correct} | \text{percept, choice})$, associated with each choice are integrated multiplicatively to generate an estimate of integrated value, $V = R \times L$. (4) Rats chose the option with the larger value. Under this assumption, the choice probability of each option is proportional to the relative value of the available choices.

Reward size: to formalize this variable, we defined the expected relative reward size, R_A , for each option in a particular block as follows:

$$R_A = \frac{W_A}{W_{\text{control}}}$$

where W_A is water reward amount for choice option A (that is, left or right) in reward-biased blocks and W_A is water reward amount in control block (W_{control}). Animals showed approximately linear dependence on relative reward size during the task.

Outcome likelihood: animals typically showed a sigmoid psychometric response profile across a range of odour cues. We estimated the

subjective likelihood, L_L , of outcome for the left choice for a particular stimulus based on the psychometric function itself. For example, if an animal responded to odour pair A/B = (55%/45%) by choosing the left port on 60% of trials, then the animal's estimated probability of success (L_L) on that odour pair would be 0.6. Note that we used the estimated outcome likelihood to explain choice probability under reward size manipulation (or previous outcome effect).

Choice value: the value of each choice was estimated as a product of outcome likelihood and relative reward size:

$$V_X = L_X (R_X w)$$

where w is the weighting coefficient that determines reward size sensitivity and X denotes the choice side (that is, left or right).

Choice probability: Choice probability for choosing left was proportional to the relative reward value:

$$P_L = \frac{V_L}{V_L + V_R}$$

and was matched to the actual behavioural choice probability function (Fig. 1d) by fitting one free parameter, w . Note that w was set to 1 for the choices associated with normal reward size (so w does not cancel out in the function above) and therefore values for left and right choices are always asymmetric in the bias block. The coefficient of determination R^2 was calculated by squaring the correlation between the model-derived estimate of P_L and actual choice data.

Previous outcome bias: we extended this model to explain choice bias following correct or error trials in the control block. Since rats could differentially weigh the value of previously rewarded and unrewarded trials (beyond the sign difference), we independently estimated subjective values after previous correct and error choices. To estimate subjective value for previously correct and error trials, we used $V_X = L_X w_C$, for previously correct and $V_X = L_X w_E$, for previously error trials. For the unchosen side, we set $V_X = L_X$ as a reference. L_X represents the outcome likelihood and X denotes the choice side (that is, left or right). Then, we fit the choice probability as a function of relative integrated value as described above to find the w_C and w_E that minimized the squared error. See Extended Data Fig. 1h–l.

As noted above, the linearized odour stimulus representation is reported using the axis legend “% odour A”.

Behavioural strategies

In the reward-biased psychometric task used here, animals' performance is limited by two factors: (1) the difficulty of the psychometric task, which produces probabilistic rewards on many trials; and (2) the reward bias across trials, which provides inferior rewards in response to some choices. Within the limits of an animal's psychometric performance, the optimal reward-maximization strategy is to generate choices based on an integrated estimate of choice value—estimated here as a multiplicative integration of outcome likelihood (confidence) and reward size associated with each choice.

To demonstrate that rats do indeed perform such optimal integration, we first modelled several inferior strategies (Extended Data Fig. 1f, g). **Psychometric-only model:** we modelled a strategy that relies on baseline psychometric performance to generate an estimate of the total reward available if an animal's choices only depended on outcome likelihood (ignoring reward size information). **Reward-only model:** we generated a model in which animals always choose the port associated with the larger reward (ignoring sensory evidence). **Integrative model:** for each session, expected reward amount was calculated as:

$$\text{Expected reward amount} = \sum_{n=1}^T (L_n R_n)$$

where L_n is reward likelihood for trial n , R_n is reward size in trial n . Reward likelihood was approximated based on the psychometric function across sessions, and T is total number of trials in a session. For psychometric-only and reward-only models, either R_n or L_n was fixed to the average values for appropriate to any given block.

Extended Data Fig. 1f, g shows that animals outperformed either the psychometric-only or the reward-only model. In addition, they showed hallmarks of reward-biases psychometric performance as described above, including the observations that the magnitude of reward bias is highest under maximal uncertainty (Fig. 1c), and that the choice probability is strongly correlated with the model (Fig. 1e). On the basis of this we infer that rats are performing a reward-maximizing integration of sensory evidence and reward contingencies, with their overall performance limited by the available sensory evidence within the difficult psychometric task.

Spike detection and sorting

Spikes were manually sorted into single-unit clusters (presumptive single neurons) off-line based on peak amplitude and waveform energy using the MClust software (A. D. Redish). Clusters were considered as single units only when the following criteria were met: (1) refractory period violations were less than 1% of all spikes; and (2) the isolation distance, which was estimated as the distance from the centre of identified cluster to the nearest cluster on the basis of the Mahalanobis distance, was greater than 20^{36} . Units were sorted blind to any other criteria.

Optogenetic identification of tagged units

To identify photo-tagged neurons we used the Stimulus-Associated spike Latency Test (SALT; see supplementary note 1 of ref. ²⁸ for a detailed description). SALT is a statistical test to determine whether optogenetic activation caused a significant change in the timing of spikes after stimulation onset. Specifically, the distribution of first spike latencies relative to the light pulse, assessed in a 10 ms window after light stimulation, was compared to epochs of the same duration in the stimulus-free baseline period. The choice of 10 ms window size provided sufficient statistical power without limiting the number of detected neurons.

Light-evoked spikes were defined based on the peak of PSTH during the 10 ms after light onset. Well-isolated single units (see above for the criterion) with significant correlation of average waveforms between spontaneous and the light-evoked spikes (Pearson's $r > 0.85$) as well as reliable spike generation (probability > 0.4 , $P < 0.01$, SALT) were identified as striatal projecting neurons. The activated neurons formed a distinct cluster in the space defined by spike latency, jitter and probability of the first spike (Extended Data Fig. 10g). Identified neurons showed strong correlation of the light-evoked and spontaneous spike waveforms (median correlation coefficient, 0.99; s.e., 0.0068; range, 0.89 to 1.0; Extended Data Fig. 10f). The median reliability of light-evoked responses was 0.61 (s.e., 0.036; range, 0.1 to 0.9) for low-frequency stimulation (10 Hz for $n = 22$ neurons; 20 Hz for $n = 2$ neurons).

Spike train analysis

We analysed single unit data from 146 behavioural sessions from 9 rats. Unless otherwise stated, spike trains were smoothed by convolution with a Gaussian kernel ($\sigma = 15$ ms) to obtain a spike density function (SDF) for the analysis of the temporal profile of neuronal activity. For most analyses, we focused our analysis on the 'reward anticipation period' while rats remained at one of the choice ports.

For cohort 1 (rats C051, C052 and C068), neuronal responses were analysed for 485 well-isolated neurons that had a non-zero firing rate during the anticipation window. For cohort 2 (rats C091, S11, V03 and V05), neuronal responses were analysed for 639 well-isolated neurons that had a non-zero firing rate during the anticipation window. For cohort 3 (rats HA56 and HA58), neuronal responses were analysed for 383 well-isolated neurons. This cohort was used only for OFC to

striatum projection neurons and because the rats were tested without reward size manipulation these were not included in any of the clustering analyses.

For cohort 1 and 2, response profiles were generated across 48 conditions (Fig. 2b) with six conditions dropped due to frequent missing values (Extended Data Fig. 5i, j). The matrix of response profiles was subject to de-noising and imputation of sparse missing values using probabilistic principal components analysis (pPCA). To retain full population diversity in downstream analyses, coefficients were truncated to retain 90% of between-neuron variance, resulting in a set of coefficients for each neuron in a 21-dimensional response space (Extended Data Fig. 3b).

Although PCA is notoriously susceptible to over-interpretation, we note that approximately the first seven (accounting for ~60% of observed variance) resemble somewhat distorted mixtures of common decision variables (the first three are shown as Extended Data Fig. 3d). The interpretability of such eigenvectors as dominant neuronal tuning curves is limited by the assumption that the data are distributed as a multivariate Gaussian, and the number of eigenvectors resembling decision variables can be smaller than the actual number of unique decision variables if those decision variables are not linearly independent. For example, a 2D space for which the basis vectors represent confidence and reward size could contain representations of three distinct decision variables (confidence, reward size and choice value). We therefore sought to understand whether neurons in such a space show random mixed selectivity, or instead form categorical representations that align to putative decision variables using more detailed analyses.

Tests for random mixed selectivity

The assumption of random mixed selectivity is pervasive in the interpretation of cortical areas, but is rarely quantified directly. Here, we outline an improved set of tests that are appropriate for multi-dimensional neuronal data, drawing on the idea that random mixed selectivity is observed when neuronal responses can be represented as a multiple Gaussian (or similar) distribution.

Random mixed selectivity in this population was assessed using two novel methods: the 'elliptical projection angle index of response similarity' test (ePAIRS) and the 'elliptical random projection' test (eRP). In brief, these non-parametric tests assess whether the set of neuronal response profile are evenly distributed throughout the representational space, while accounting for differences in dimension size and tolerating changes in the magnitude of each response.

We formulate our approach as follows. We will first recapitulate a fairly standard approach to analysis of neuronal populations, and then discuss a deviation from some of its assumptions. We assume that we have recorded a set of n q -dimensional response profiles $X^{n,q}$. Each column of X corresponds to a single stimulus or task condition, and each row corresponds to a single neuron. This set of response profiles is usually z-scored, and de-noised to d dimensions using principal components analysis, retaining only a truncated set of d -dimensional coefficients for each neuron, along with a set of d q -dimensional orthogonal loading vectors.

For a set of z-scored response profiles as a row of X , PCA follows as $X'X = U\Sigma U'$ where columns of U encode the set of n loading vectors, and scores or coefficients are recovered as $S = XU$. The loading vectors are ranked by their contribution to the overall between-neuron variance of the neuronal population, represented by an associated eigenvalue on the diagonal of Σ . For the truncated decomposition using only the first d columns of U , S can be considered a de-noised rotation of X . As such, the first columns of B represent the axes of maximum population variance.

PCA has an intrinsic assumption of normality, and a stronger assumption that each dimension of X is independent. If this is met, loadings of S follow a multivariate Gaussian distribution, with the size of each dimension representing the between-neuron variance in the underlying variation.

Under the assumption of normality, the coefficients in S for each neuron are drawn independently for each neuron—if (and only if) this is true of the population as a whole, it can be said to show random mixed selectivity.

As an example, we present the population of synthetic neurons in Extended Data Fig. 4a–d. In each panel we show the coefficients of neurons in two dimensions; these neurons are also split into two subpopulations (blue and red). The population in the top panels (Extended Data Fig. 4a, b) have equal variance in each dimension, where those in the bottom panels (Extended Data Fig. 4c, d) have unequal variance. All of these populations may show simple mixed selectivity as long as the dimensions of variance (that is, loading vectors contained in columns of U) arise as combinations of relevant task variables (labelled here as Dimension 1 and Dimension 2).

However, this is not equivalent to random mixed selectivity, which is a statement of how the representation of each dimension is distributed across the population. The populations shown in the left panels (Extended Data Fig. 4a, c) show independent coefficients for each dimension, and thus show random mixed selectivity.

Neurons in the right panels (Extended Data Fig. 4b, d) have the same overall variance structure across the entire population: that is, the size of each dimension in the right panels matches the size of each dimension in the left panels. However, they do not show random mixed selectivity, because each neuron belongs to a subpopulation in which the coefficients in each dimension are no longer independent.

Many tests of random mixed selectivity begin from the intuition of symmetry: if the coefficients of a distribution are drawn independently and identically distributed (i.i.d), then the overall distribution should be rotation invariant, and show spherical symmetry²³. This intuition is correct in the case where dimensions are of equal size (that is, spherical distributions; Extended Data Fig. 4a, b), but not where dimensions are of unequal size (that is, elliptical distributions; Extended Data Fig. 4c, d). Notably, essentially all neuronal datasets are elliptical for the simple reason that some representations are more prevalent than others.

Here we adapt and validate two tests for spherical symmetry that can be used for elliptical distributions. We define these tests based on two recently reported non-parametric tests for rotational invariance that are applicable to spherical distributions: the projection angle index of response similarity test (PAIRS²²), and a modified random projection test (RP²³).

The PAIRS test was first presented by ref.²². PAIRS approaches the problem as follows. (1) Given a d -dimensional dataset $X(n \times d)$, calculate for each row in X the cosine distance to its nearest neighbour. Calculate the empirical median of these nearest-neighbour distances as \tilde{e} . (2) Generate a set of m bootstrap distributions $Y(Y_1, \dots, Y_m)$, as spherical d -dimensional Gaussians. For each Y^* , calculate the distribution of nearest-neighbour angles, and pool these measurements to generate a bootstrap distribution $B(b_1, \dots, b_m)$. (3) Calculate the likelihood that the dataset X shows significant clustering as the empirical expectation $P = E(\tilde{e} < b)$.

The intuition here is that, for any plausible form of clustered data, median nearest neighbour distances are likely to be smaller than those for uniformly distributed data. This assumption does not hold under some pathological but non-random distributions (for example, points that are regularly spaced).

The RP test was first presented by ref.²³ as the RPK test, with our implementation based on that of B. Lau (<https://github.com/brian-lau/highdim>). The RP test is similar to PAIRS in that it derives from a comparison of angle distributions. Specifically, for a spherical d -dimensional dataset X , we (1) generate a set of k random vectors as rows of a matrix $Z(k \times d)$, drawn uniformly across d dimensions. By default, the dimensions are of equal size (but see below). (2) Calculate the distribution of projection angles for each row in Z onto each row in X . Each of these k distributions is notated as a vector e_k . (3) Generate a set of m bootstrap distributions $Y(y_1, \dots, y_m)$, as spherical d -dimensional

Gaussians. Calculate the distribution of projection angles for each point in each distribution Y^* onto each vector in Z . Each such distribution is notated as $B_{m,k}$. (4) For each k , calculate the Kolmogorov–Smirnov statistic between the empirical distribution e_k and each bootstrap sample $B^*_{m,k}$. The median Kolmogorov–Smirnov statistic is designated as the empirical observation \tilde{e} . (5) For each k , calculate Kolmogorov–Smirnov statistics between all $m(m-1)$ pairs $B_{m,k}$. Pool all such Kolmogorov–Smirnov statistics as the bootstrap distribution b . (6) Compare the median empirical-versus-bootstrap observation \tilde{e} to the bootstrap-versus-bootstrap distribution b . As larger Kolmogorov–Smirnov statistics reflect larger deviations between distributions, we define the probability that X is spherically uniform as $P = E(\tilde{e} < b)$.

We note that this formulation is distinct from the RPK test proposed by Cuesta–Albertos and colleagues²³, who suggested a direct comparison of Kolmogorov–Smirnov P values with Bonferroni correction. We find that method to have low specificity and to be unstable under minor noise levels, and suggest the Monte Carlo procedure used here as a more reliable alternative.

As validation of our concern about sphericity, we note that both PAIRS and RP will correctly categorize the spherical uniformity of Extended Data Fig. 4a, and correctly detect the deviation from spherical uniformity in Extended Data Fig. 4b, d. However, both tests will also report that the distribution in Extended Data Fig. 4c is not spherically uniform (Extended Data Fig. 4e, i). This is at odds with the intuition that the elliptical distribution in Extended Data Fig. 4c still shows random mixed selectivity.

As a compensation for this effect, we can simply alter both the RP test and PAIRS test to use elliptical bootstrap distributions that match the variance structure of the test data. This is done by drawing bootstrap distributions Y from elliptical normal distributions that match the distribution sizes λ_i observed in X .

$$Y = N(0, 1, D)$$

$$Y = \{y_i \times \lambda_i\} \text{ for each dimension size } \lambda$$

The resulting set of vectors can be used to generate a bootstrap distribution, because although it is not uniform on the unit hypersphere, it does arise from a matching spherically uniform Gaussian distribution. We refer to the resulting tests as ‘elliptical’ tests: ePAIRS and eRP.

To validate our approach, we conducted simulations using a set of point distributions that matched the variance structure observed in our OFC data (Extended Data Fig. 3c), but were otherwise either distributed uniformly or contained synthetic clusters generated as von Mises distributions (Extended Data Fig. 4e, f) and eRP (Extended Data Fig. 4i, j). PAIRS and ePAIRS tests were conducted with 1,000 bootstrap samples, across 30 replicates. The RP and eRP tests were conducted with 20 bootstrap distributions and 500 samples in each bootstrap distribution.

As expected, we observed that both the spherical PAIRS and RP tests produced frequent false-positive results for data that were known to be drawn from a uniform elliptical distribution (Extended Data Fig. 4e, i). The improved eRP and ePAIRS tests, however, generally correctly distinguish between clustered and un-clustered data (Extended Data Fig. 4f, j). Generally, the eRP test appears to have better sensitivity in identifying non-uniformity in the form of von Mises distributions, but we have not attempted to generalize this result to other models of non-uniformity.

This approach gives rise to several empirical results. First, we observe that the spherical RP and PAIRS tests are subject to extremely high type 1 (false positive) errors when presented with data from an elliptical Gaussian distribution. Nevertheless, this result does not necessarily invalidate the approach of ref.²², for which the non-corrected test is simply the most conservative case for the result they ultimately observed (that is, apparent uniformity and random mixed selectivity).

After correction, we observe that the elliptical random projection (eRP) and elliptical PAIRS (ePAIRS) tests correctly fail to reject the hypothesis of spherical uniformity for the uniform Gaussian distribution, but correctly detect the deviation for the clustered von Mises distributions. In general, the eRP test appears to have greater sensitivity for weakly clustered von Mises distributions.

An implicit variable in much of this analysis is the dimensionality of the observed dataset \mathbf{d} . Our approach throughout this paper has been to only minimally de-noise our dataset and retain 21 dimensions that account for 90% of population variance. However, analyses presented in Extended Data Fig. 4e, f, i, j account for variations in \mathbf{d} and show that our main result is invariant across \mathbf{d} .

Spectral clustering

Spectral clustering was performed using standard techniques on a binary adjacency matrix^{37,38}, including Shi and Malik normalization of the Laplacian followed by k-means clustering of the small eigenvectors (Extended Data Fig. 3f).

The adjacency matrix was generated by identifying the k-nearest neighbours under a cosine or correlation distance metric. Hyperparameters, including number of nearest neighbours (k), the number of clusters (c), were selected by maximizing the adjusted Rand Index (ARI). For ARI analysis, 100 bootstrap samples were generated for each hyperparameter combination, with a sub-sampling fraction of 0.9. Similar results were observed for both cosine and correlation distance metrics; the results of a correlation metric are shown in Fig. 2e, f. Peak ARI in this dataset (ARI = 0.65) was significantly higher than expected from trial-shuffled datasets (ARI = 0; $P < 0.01$; t -test), or from datasets in which each dimension of the response profile was independently shuffled (ARI = 0; $P < 0.01$; t -test).

Regression analysis

For the decision-variable model, we first defined a set of five putative decision variables: (1) the choice that the rat just made (left or right); (2) stimulus uncertainty (that is, stimulus difficulty conditional on choice); (3) expected reward size for each trial; (4) outcome of the previous trial; and (5) overall block type (biased or unbiased). Each variable was parameterized as a z-scored vector corresponding to the conditions in which it was relevant, with irrelevant conditions masked by setting them to zero. This set of vectors—appropriately masked and z-scored—was used as the model or design matrix for the regressions that follow.

Using this design matrix, we fit the z-scored response profile of each neuron, or the z-scored average response profile arising from each cluster of neurons. Because some conditions may be missing for each neuron, these were imputed using probabilistic PCA (see ‘Spike train analysis’).

For the regression in Fig. 4c, d, these variables were used in a non-penalized linear regression. For the penalized regression in Fig. 4e–h, these variables were used in a LASSO regression using the glmnet package. We used the traditional LASSO representation

$$\min_x \frac{1}{2} \|y - (Ax + x_0)\|^2 + \lambda \|x\|_1$$

where y is the response profile of a neuron or cluster, A is the design matrix, and x is the set of recovered coefficients.

The regularization parameter λ was selected to minimize the error of tenfold cross-validation. The overall error of each fit was measured as root-mean square error (RMSE).

To generate each of the null models, we applied a random rotation of the design matrix, equivalent to rotating the axes of the LASSO cartoon depicted in Fig. 4e. In a non-penalized regression, this rotation results in no change in overall fitting error. In an L1-penalized regression, however, the L1 penalty will rise, but only if the output

variable (that is, the neuronal response profile) does in fact represent a sparse combination of the elements of the design matrix (that is, the decision variables).

For each regression, we computed the reconstruction error and mixture penalty (L1 penalty, or sum of absolute value of all coefficients). We observed significant increase in the mixture penalty when the regression was performed with the non-canonical null models. The average values for RMSE and L1 penalty are shown in Extended Data Fig. 8g.

Note that in Fig. 4f, g we omitted the cluster corresponding to ‘conditional updating’ for this analysis, because all of the neurons in this cluster arose from a single rat and because the response profile of this cluster was a poor fit to any hypothesized set of decision variables.

We performed the same regression on the response profiles of individual neurons. For some conditions, we had relatively few trials to enter into the average response (<5); therefore, we limited our approach to a twofold cross-validation, repeated 20 times. We examined the RMSE and L1 penalty of fits using the canonical decision variable model as well as a set of 20 random rotations of that model. In total, we performed 400 regressions for each of 485 neurons, scanning across a set of 10,000 lambda values. We observed that the L1 penalty was significantly lower for regressions using the canonical model as opposed to rotated models ($P = 0.002$; sign rank test; Extended Data Fig. 8f) with no difference in overall error.

In addition, we repeated this analysis after removing neurons that were poorly fit by the regression (that is, lacking a local minimum in reconstruction error, or for which all regression coefficients were less than 1×10^{-3}) and observed similar results (retained 319 cells; $P = 0.01$, sign rank test).

To provide a complementary analysis that did not rely on cross-validation, we performed a similar regression in which the neuronal response profile of all neurons was fit via a multi-response Gaussian with no penalization. This approach minimizes the Frobenius norm of the full error matrix:

$$\min_x \frac{1}{2} \sum \|y - (Ax + x_0)\|_F^2$$

which has the convenient property that it is invariant under rotation (that is, simultaneous rotation of the neuronal response matrix and the design matrix).

This analysis produces 2,425 coefficients total (arising from 485 neurons and 5 coefficients per neuron). We treated this as a single distribution, and considered that the kurtosis of this distribution would be highest when the decision variables contained in the design matrix were maximally aligned with individual neuronal responses (Fig. 4g). This intuition is analogous to that underpinning independent components analysis, which directly seeks the axes on which data deviate from a Gaussian distribution³⁹.

We therefore compared the kurtosis of the coefficients distribution under the canonical decision variable model to that observed for fits under a set of random rotations (Fig. 4h). As expected, we observed significantly higher kurtosis under the canonical model than expected from random rotations ($P = 0.000266$, sign rank test; $P < 0.002$, bootstrap test). Together with the results of the standard LASSO analysis, these results demonstrate that individual neuronal response profiles are sparsely loaded onto the decision variable model.

Epoch analysis

We repeated the same clustering procedure as described above for two additional task epochs (the stimulus epoch and the feedback epoch). For the stimulus epoch, we included all spikes from the start of odour delivery until the rat left the centre port. For the feedback epoch, we included all spikes in the 500 ms following the revelation of the outcome. See Supplementary Note 3 ‘Epoch analysis’ for further details.

Temporal dynamics clustering

For temporal clustering, we first constructed a feature vector for each neuron by calculating the average firing rate in 10 ms consecutive, non-overlapping time bins smoothed with a Gaussian kernel ($\sigma = 50$ ms) and aligned to several task events (that is, centre poke, stimulus delivery, side poke, outcome). To allow for averaging firing rates across trials despite variable task epochs, we divided each task epoch into a fixed number of bins (one bin on average corresponds to 10 ms) and calculated the firing rate for each bin (pre-poke epoch, 50 bins; pre-stimulus epoch, 45 bins; stimulus epoch, 50 bins; movement epoch, 30 bins; anticipation epoch, 100 bins; feedback epoch, 150 bins; see figure 10 in ref.⁴⁰). We then averaged the firing rate for each bin across trials to construct a trial-averaged, time-normalized feature vector separately for all correct and all error trials. Next we generated the full feature matrix by aligning the z-scored firing rates of all recorded neurons (cohort 1 and Cohort 2 with at least 1 Hz firing rate and sufficiently long pre-stimulus period), concatenating correct and error trial averages to yield a 850-dimensional feature matrix with $n = 679$ neurons. We reduced the number of feature dimensions (that is, time points) using PCA retaining 90% of variance (14 PCs). This feature space was used for spectral clustering following the same procedure as above. Spectral clustering yielded robust clusters validated by bootstrap stability using the adjusted rand index and the most stable clustering was obtained for hyperparameters $n = 8$ clusters and $k = 13$ nearest neighbours.

No statistical methods were used to predetermine sample size.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author upon request.

Code availability

Software for ePAIRS and eRP is available at <https://github.com/KepecsLab/EllipticalClustering>.

31. Sanders, J. I. & Kepecs, A. A low-cost programmable pulse generator for physiology and behavior. *Front. Neuroeng.* **7**, 43 (2014).
32. Gradinaru, V. et al. Molecular and cellular approaches for diversifying and extending optogenetics. *Cell* **141**, 154–165 (2010).
33. Li, S. J., Vaughan, A., Sturgill, J. F. & Kepecs, A. A viral receptor complementation strategy to overcome CAV-2 tropism for efficient retrograde targeting of neurons. *Neuron* **98**, 905–917.e5 (2018).
34. Soudais, C., Laplace-Builhe, C., Kissa, K. & Kremer, E. J. Preferential transduction of neurons by canine adenovirus vectors and their efficient retrograde transport in vivo. *FASEB J.* **15**, 2283–2285 (2001).
35. Tervo, D. G. R. et al. A designer AAV variant permits efficient retrograde access to projection neurons. *Neuron* **92**, 372–382 (2016).
36. Schmitzer-Torbert, N., Jackson, J., Henze, D., Harris, K. & Redish, A. D. Quantitative measures of cluster quality for use in extracellular recordings. *Neuroscience* **131**, 1–11 (2005).
37. Von Luxburg, U. A tutorial on spectral clustering. *Stat. Comput.* **17**, 395–416 (2007).
38. Shi, J. & Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 888–905 (2000).
39. Comon, P. Independent component analysis, A new concept? *Signal Process.* **36**, 287–314 (1994).
40. Kobak, D. et al. Demixed principal component analysis of neural population data. *eLife* **5**, e10989 (2016).

Acknowledgements We would like to thank past and present members of the Kepecs Laboratory and B. Mensh for many valuable discussions; J. Sanders and B. Hangya for help with experimental setup; and B. Mensh, T. Gouvea, M. Kaufman and A. Lak for comments on an earlier version of this paper. This study was funded by the grants from the Klingenstein, Alfred P. Sloan, Swartz, Whitehall Foundations and NIH grants R01DA038209 and R01MH097061 (A.K.), and KAKENHI 16K18380, 16H02061, 19H05028 (J.H.).

Author contributions J.H. designed and performed the experiments. J.H. and A.V. designed and performed the primary analyses, and T.O. and P.M. conducted additional analyses. A.K. designed the experiments, analyses and supervised the project. All authors contributed to writing the manuscript.

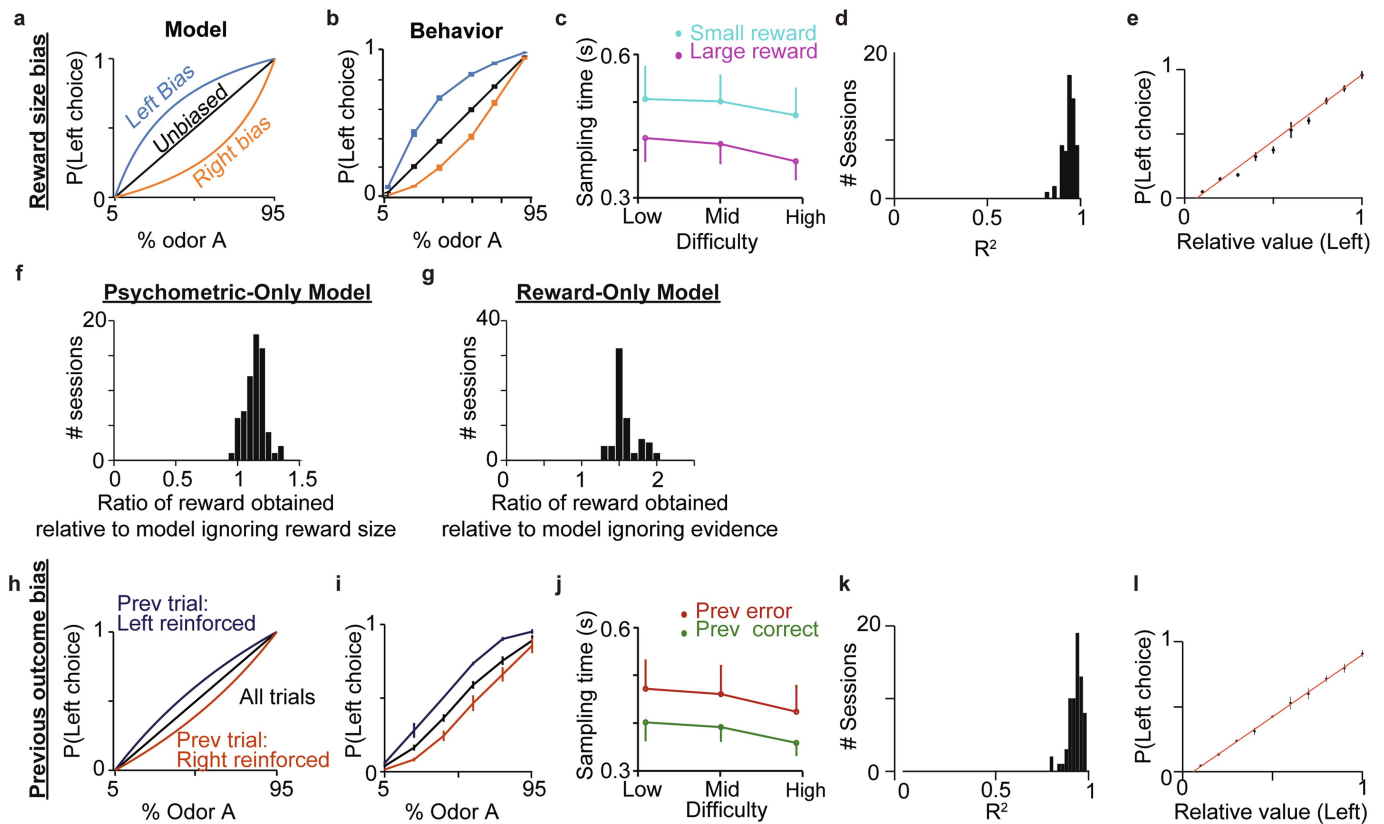
Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1816-9>.

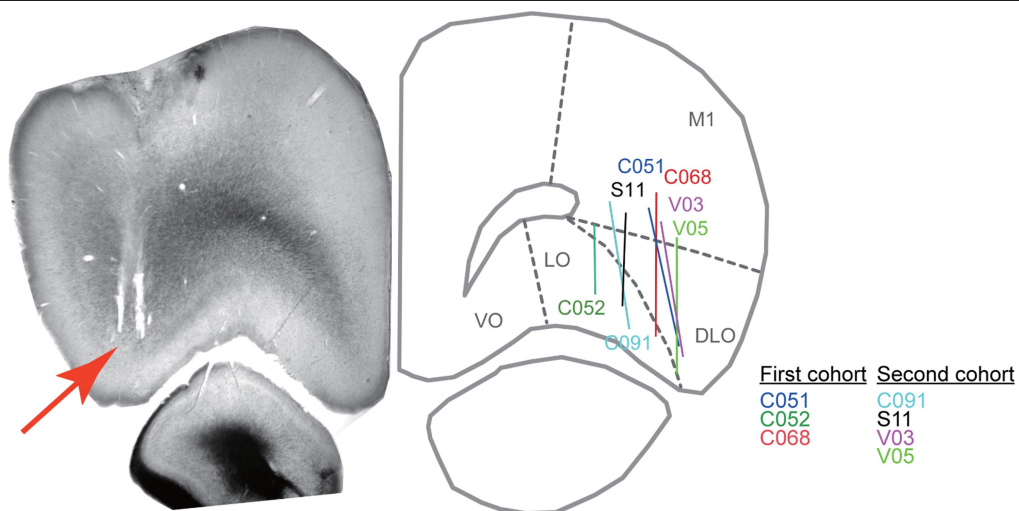
Correspondence and requests for materials should be addressed to A.K.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

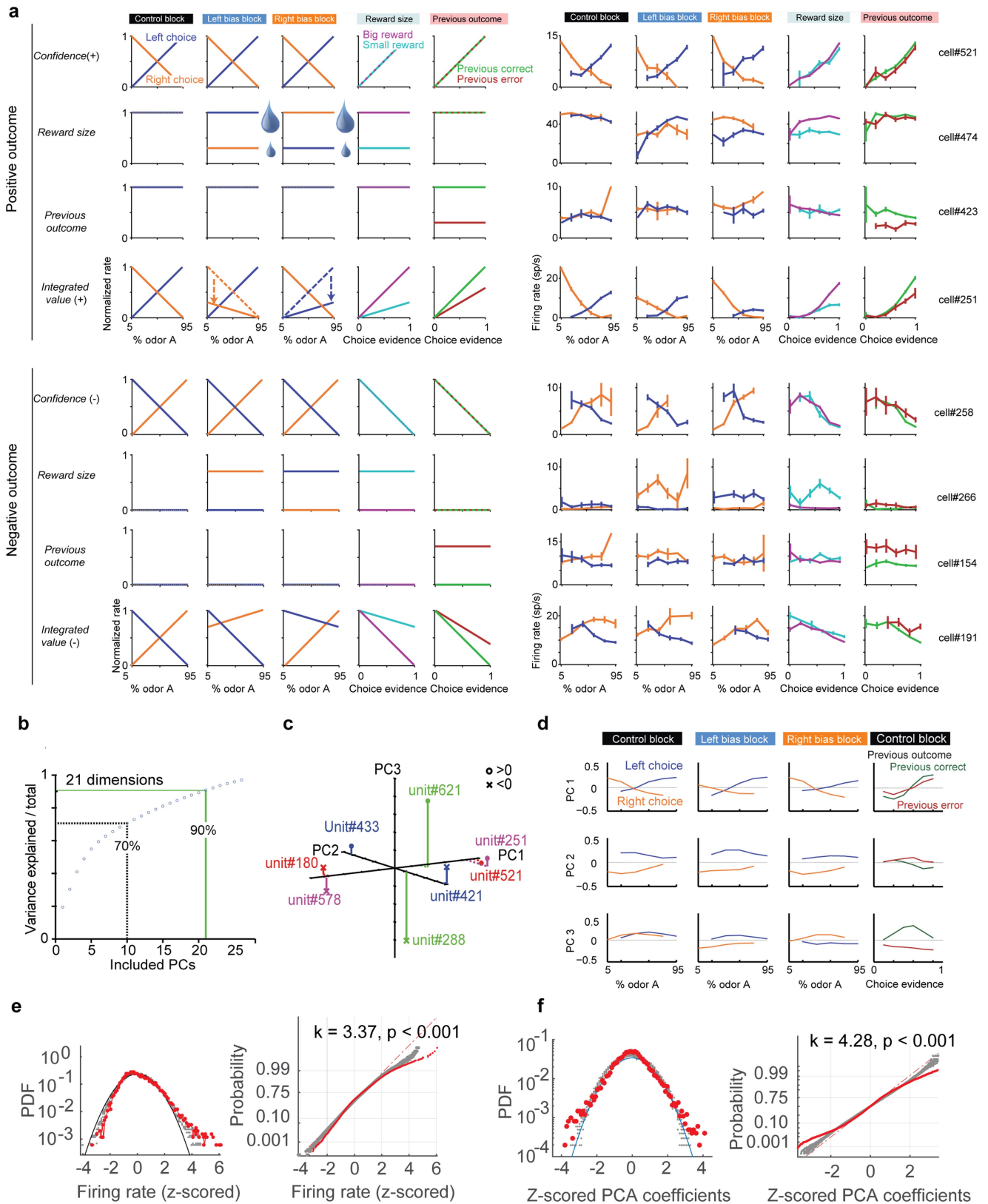


Extended Data Fig. 1 | Rat behaviour reflects an integration of evidence and reward-size. **a–e**, Average psychometric functions in unbiased, left and right bias blocks as a function of the odour percept ($n = 67$ sessions). The decision-variable model (**a**) and actual data from the same data (**b**) (replication of Fig. 1c) are shown. **c**, Odour sampling time was larger for small rewards than for large rewards. Errors are shown as mean \pm s.e.m. ($n = 3$ rats). **d**, The model provides an excellent fit of choice patterns for each session. **e**, Choice driven by the relative value of left choice (replication of Fig. 1e). **f**, Histogram across sessions, comparing the ratio of actual reward obtained to a model relying on odour stimuli but ignoring reward size. **g**, Histogram across sessions, comparing the ratio of actual reward obtained to a model relying on reward size but ignoring odour stimuli. **h–l**, Same convention as **a–e** but reporting the bias arising from

the outcome of the previous trial during the control block without reward bias ($n = 67$ sessions). **h, i**, The probability for left choice as a function of odour percept for all trials or separated by which choice (left/right) was rewarded in the previous trial. 'Left reinforced' indicates that rats are rewarded (correct) on the left side in the previous trial or not rewarded (error) on the right side, regardless of the stimulus conditions used in previous trials. The decision-variable model accurately predicts changes in choice probability (**h, i**) arising due to previous outcome. **j**, Odour sampling time was larger after a previously unrewarded choice than for a previously rewarded choice. Errors are shown as mean \pm s.e.m. ($n = 3$ rats). The model provides an accurate fit across sessions (**k**) driven by the relative value of the left choice (**l**).



Extended Data Fig. 2 | Recording sites in lateral OFC and functional clustering across rats. Recording sites in lateral OFC for all seven rats from cohort 1 and cohort 2 are shown. Histological section shown in left (rat C068), in which the red arrow indicates the tip of the tetraode bundle.

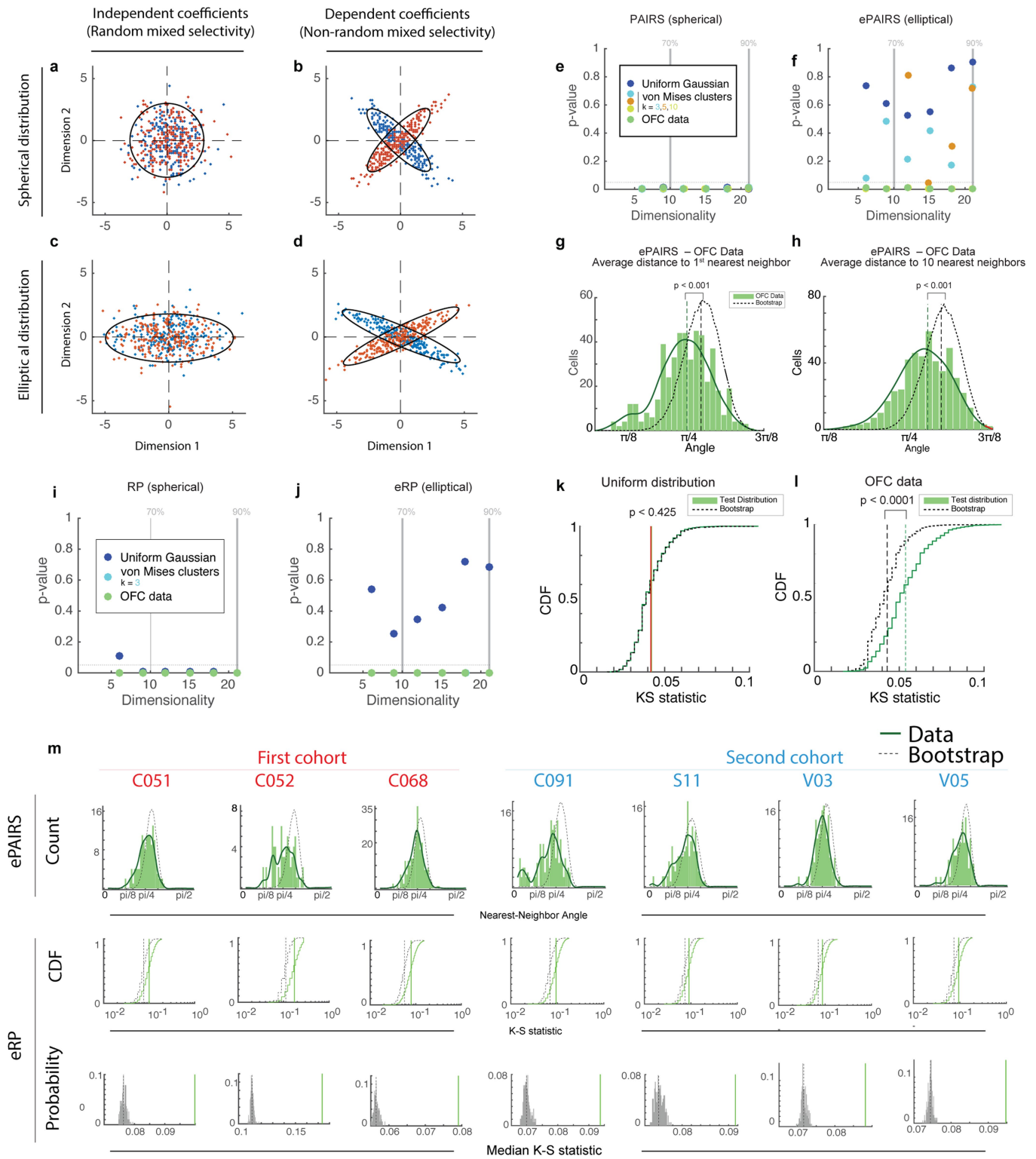


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Profiles of individual neuron responses and

population profiles in OFC. a, Example response profiles are shown for several individual neurons. In each case, a schematic tuning curve representing a plausible decision variable is shown in left panels, while a matching neuronal response profile is shown in right panels. **b,** PCA decomposition of the OFC dataset reveals high dimensionality, with 21 principal components required to retain 90% of response profile variability. **c,** Diversity of tuning vectors for several example neurons in the space of the first three principal components. **d,** The three dominant principal components arising from a probabilistic PCA

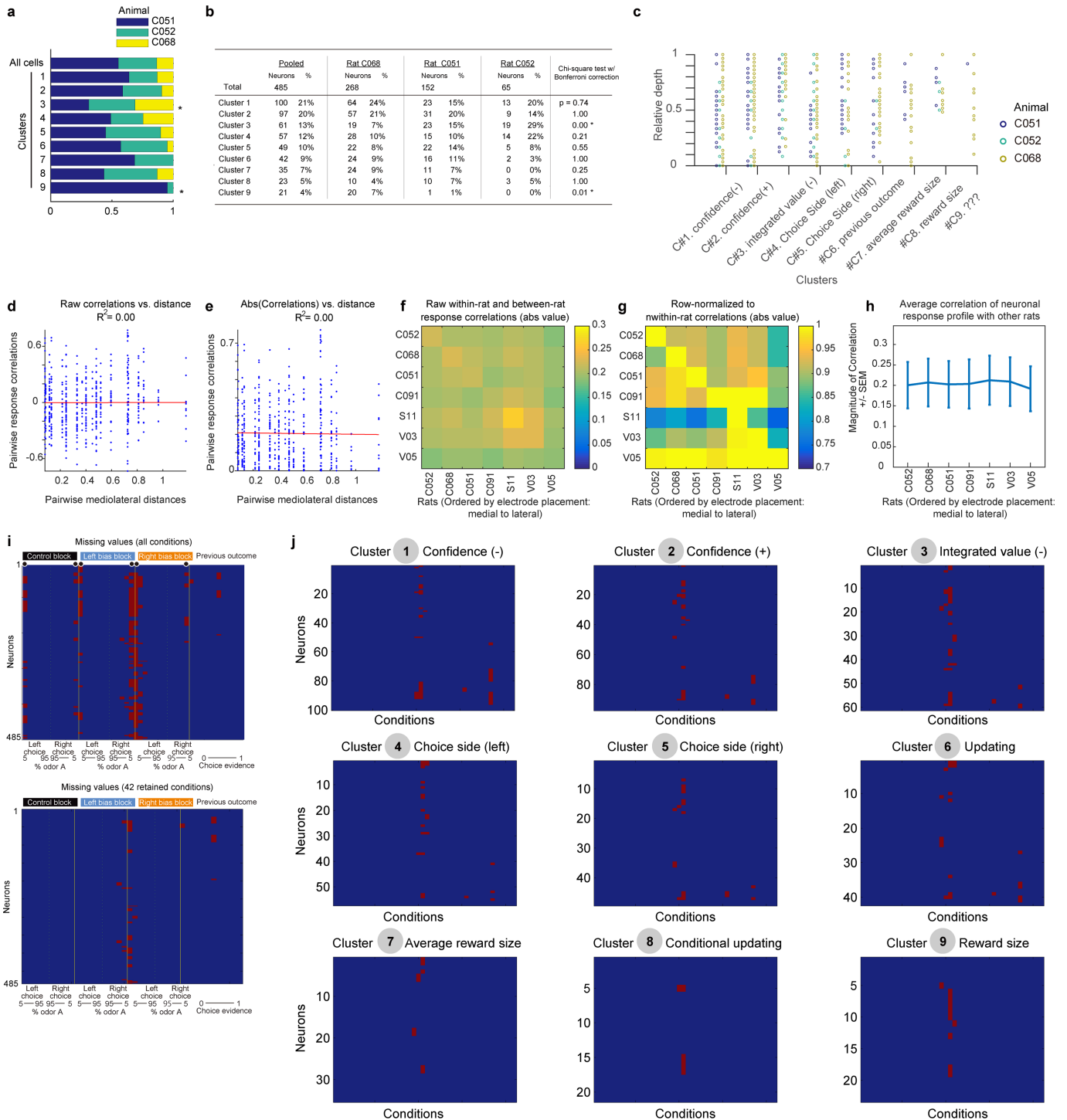
decomposition of 485 OFC response profiles, which account for ~40% of population variance. **e,** Distribution of firing rates for all neurons across all conditions. Firing rates for observed data (red dots) show a right-tailed distribution, with strong activation of most neurons for only a small subset of conditions. This pattern of activation is significantly sparser than expected from a normal distribution (black line) or trial-shuffled data (grey dots). **f,** Coefficients arising from PCA analysis show a similar long-tailed distribution compared to a normal distribution (blue line) or trail-shuffled data (grey dots).

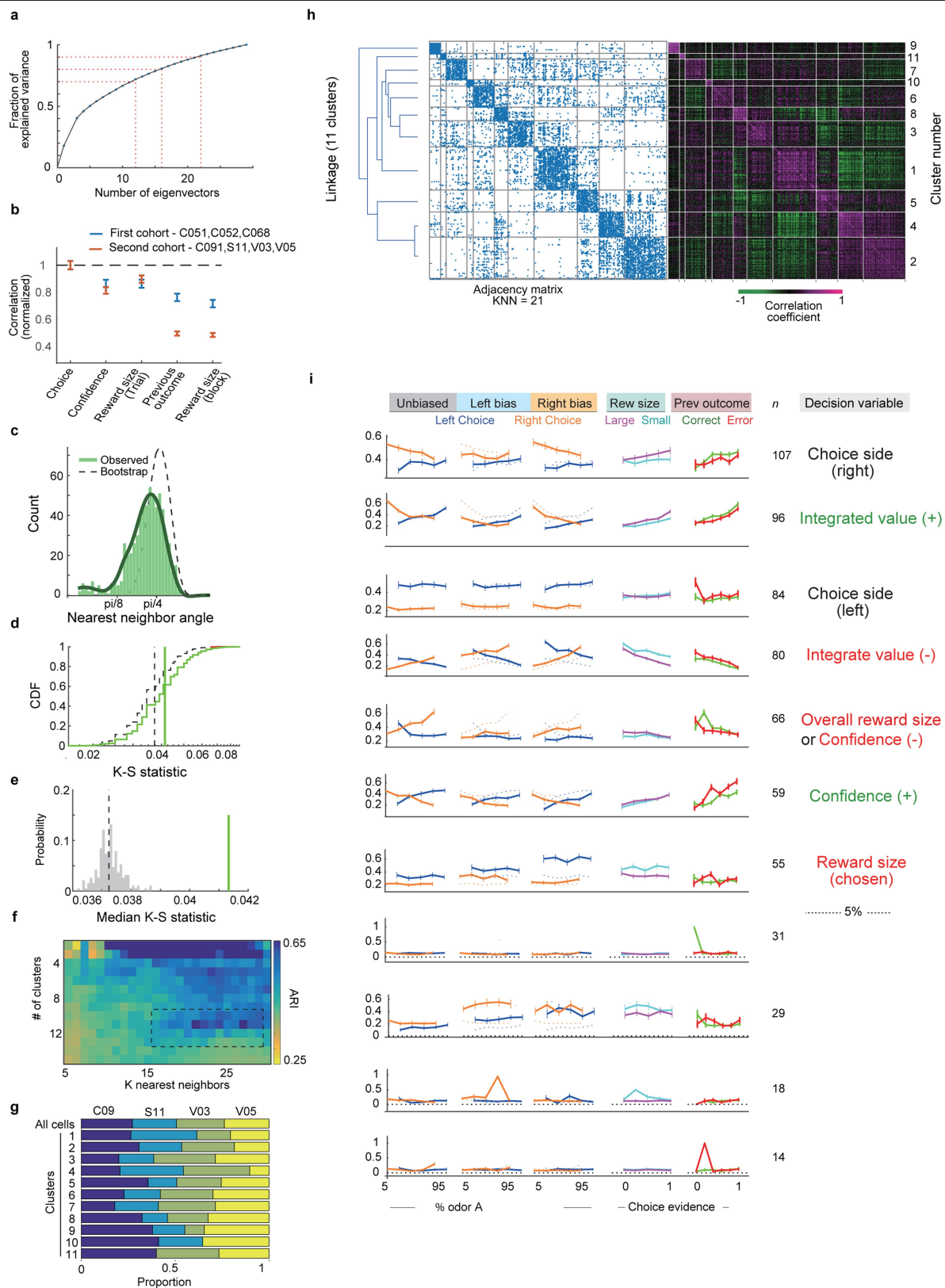


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Testing for random mixed selectivity. a–d, Exemplar distributions of toy model neurons showing mixed selectivity, presented using the coefficients for two dominant eigenvectors (dimension 1 and dimension 2). Two neuronal subpopulations are shown in red and blue. Populations in the top panels (**a, b**) have equal variance in both dimensions, whereas populations in the bottom panels (**c, d**) do not. Populations in the left panels (**a, c**) can be said to show random mixed selectivity, whereas distributions in the right panels (**b, d**) do not. Only the distribution shown in panel **a** can be said to show spherical symmetry, necessitating the development of modified tests. **e–h,** Comparison of the PAIRS to the modified ePAIRS test, which accounts for elliptical distributions. **e, f,** Sensitivity analysis of PAIRS and ePAIRS, tested across several datasets whose variance structure matched the OFC data (green), including a spherically uniform Gaussian (blue) and collections of five randomly oriented von Mises distributions with varying κ ; blue, orange, yellow). Datasets were truncated at a given dimensionality, and mean P values are reported across 30 replicates. Results show that spherical PAIRS generates false-positive results when tested on non-spherical but otherwise i.i.d Gaussian data. The modified ePAIRS test successfully identifies the non-uniformity of strongly clustered data ($\kappa = 10$) but not weaker clustering. The dimensionality required to reconstruct 70% and 90% of the variance in the full dataset is shown (grey lines). **g, h,** The ePAIRS measure, nearest-neighbour angles, is shown for OFC data (green) and bootstrap distributions (black). OFC data showed smaller angles than expected for both 1 and 10 nearest neighbours, suggesting strong clustering. **i–l,** Comparison of the random projection (RP) test to the elliptical random projection (eRP) test (eRP). **i, j,** We compared the RP and eRP tests on several datasets including an elliptical Gaussian distribution (dark blue) and

collections of five randomly oriented von Mises distributions with $\kappa = 3$ (light blue), and observed OFC data (green). Other parameters matched panels **e, f**. Results show that spherical RP generates false-positive results when tested on non-spherical but otherwise i.i.d Gaussian data (dark blue). The modified eRP test successfully identifies the non-uniformity of von Mises clusters, as well as OFC data, while rejecting spherically uniform Gaussian data. **k, l,** Cumulative distribution function (CDF) of Kolmogorov–Smirnov statistics arising in the eRP test. Results are shown for test distributions that are spherically uniform Gaussians (**k**), as well as for OFC data (**l**). **m,** Analysis of individual rats from cohort 1 and cohort 2. All rats showed significant deviation from uniformity for both ePAIRS and eRP tests. Top, histogram of nearest-neighbour angles for observed data (green) and bootstrap samples with similar elliptical distribution (black). All animals showed significant differences, assessed using a rank sum test. Middle, cumulative distribution of Kolmogorov–Smirnov test statistics from eRP for bootstrap samples and observed data. Here each observation is the K-S statistic derived from comparing the distribution of projected angles onto a single random vector between observed data and a matching elliptical Gaussian distribution. This procedure is repeated for a set of k random vectors to generate the plotted distribution. This calculation is performed both for observed data (green) and for a simulated spherically uniform bootstrap distribution with matching samples size and ellipticity (black). Bottom, comparison of the observed median K-S statistic from observed data (green), to the distribution of medians observed across several realizations of a bootstrap distribution (grey). For all eRP estimates, higher K-S statistics denote greater deviation from uniform distribution, and all rats showed significant differences using a bootstrap test.



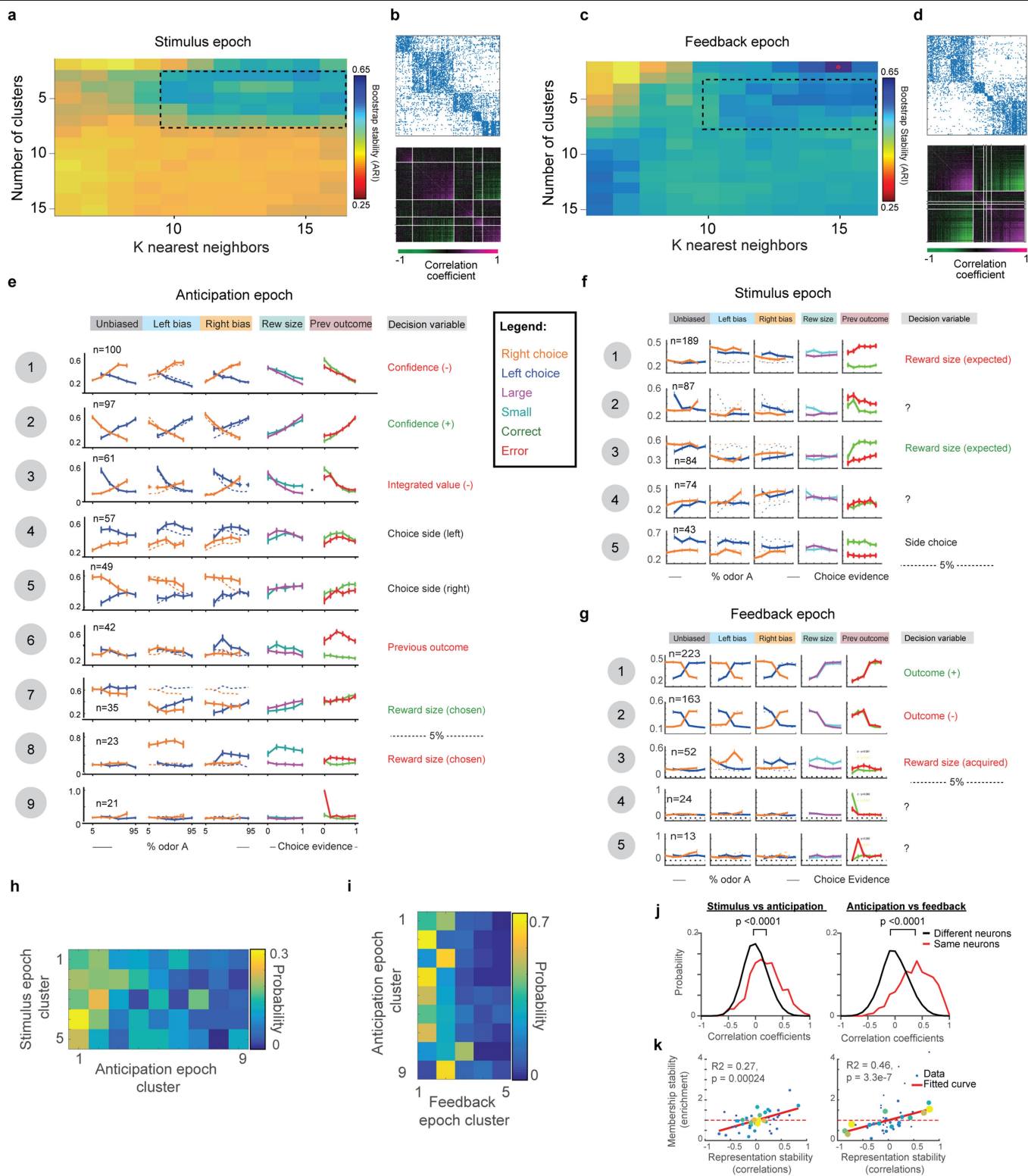


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Response profiles of OFC neurons from cohort 2

replicate the results of cohort 1. a, Distribution of variance across the first 29 eigenvectors for cohort 2, see Extended Data Fig. 3b for corresponding panel for cohort 1. **b,** Average correlations of individual cell response profiles with a set of canonical response profiles corresponding to decision variables side choice, confidence, reward size (trial-by-trial), previous outcome, and reward size (block average). The sign of the correlation was discarded and normalized across cells by the strongest correlation (that is, side choice). Two representations (previous outcome and block-wise reward size) showed reduced representation in animals from cohort 2. **c,** ePAIRS test, showing the distribution of nearest-neighbour distances between observed data in cohort 2 (green) and a bootstrap distribution derived from simulated data with a matching elliptical Gaussian (black). $P < 0.001$, Rank sum test. **d,** Cumulative distribution function (CDF) for observed data (green) and a bootstrap

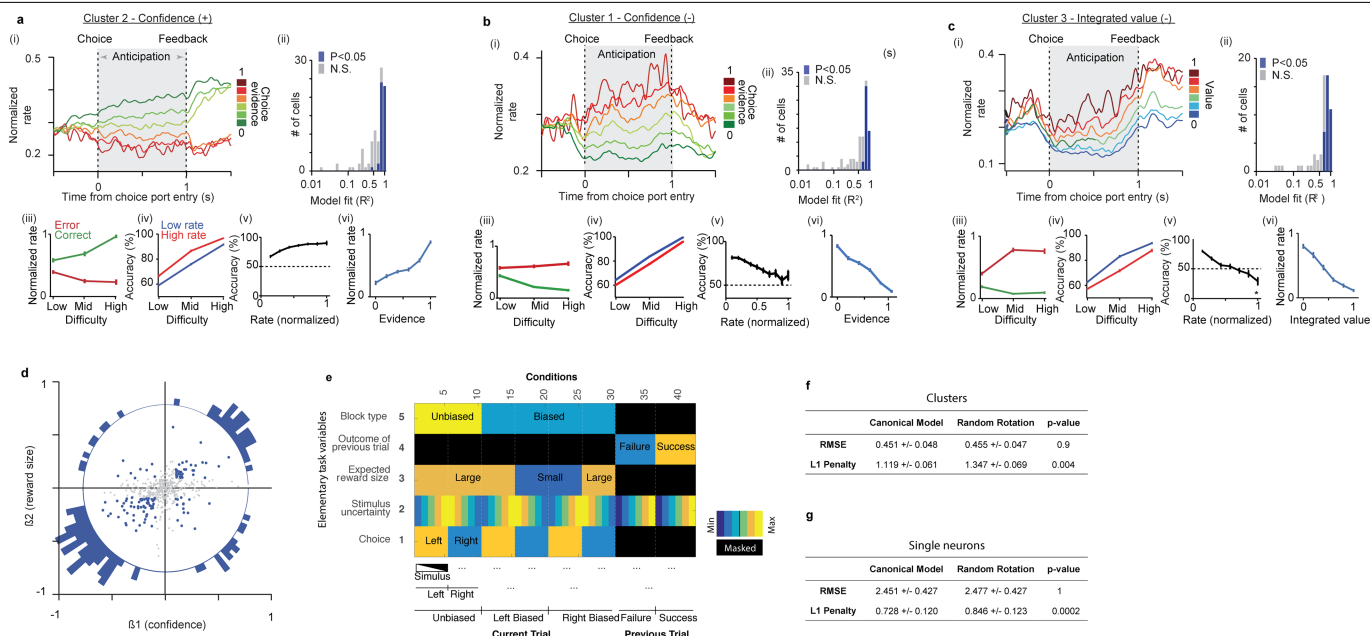
distribution (black). Median values are shown with vertical lines. **e,** Distribution of median values for the K-S statistic across a set of bootstrap distributions (grey), compared to the median value for observed data (green). $P < 0.001$, bootstrap test. **f,** ARI for spectral clustering across k (k -nearest neighbours used to generate the adjacency matrix) and number of clusters, showing marked clustering around 11 clusters. **g,** Proportion of cells from each animal associated with each cluster. **h,** Left, dendrogram of inter-cluster distances. Middle, adjacency matrix, derived from $k = 21$ using a correlation distance. Right, between- and within-cluster correlations. **i,** Average response profiles for each cluster in cohort 2. The format of this figure matches Fig. 3 in the main text. Overall, we identified 11 clusters, of which the top 7 (each containing $>5\%$ of the cells in the dataset) correspond to separable representations of choice, confidence and value.



Extended Data Fig. 7 | See next page for caption.

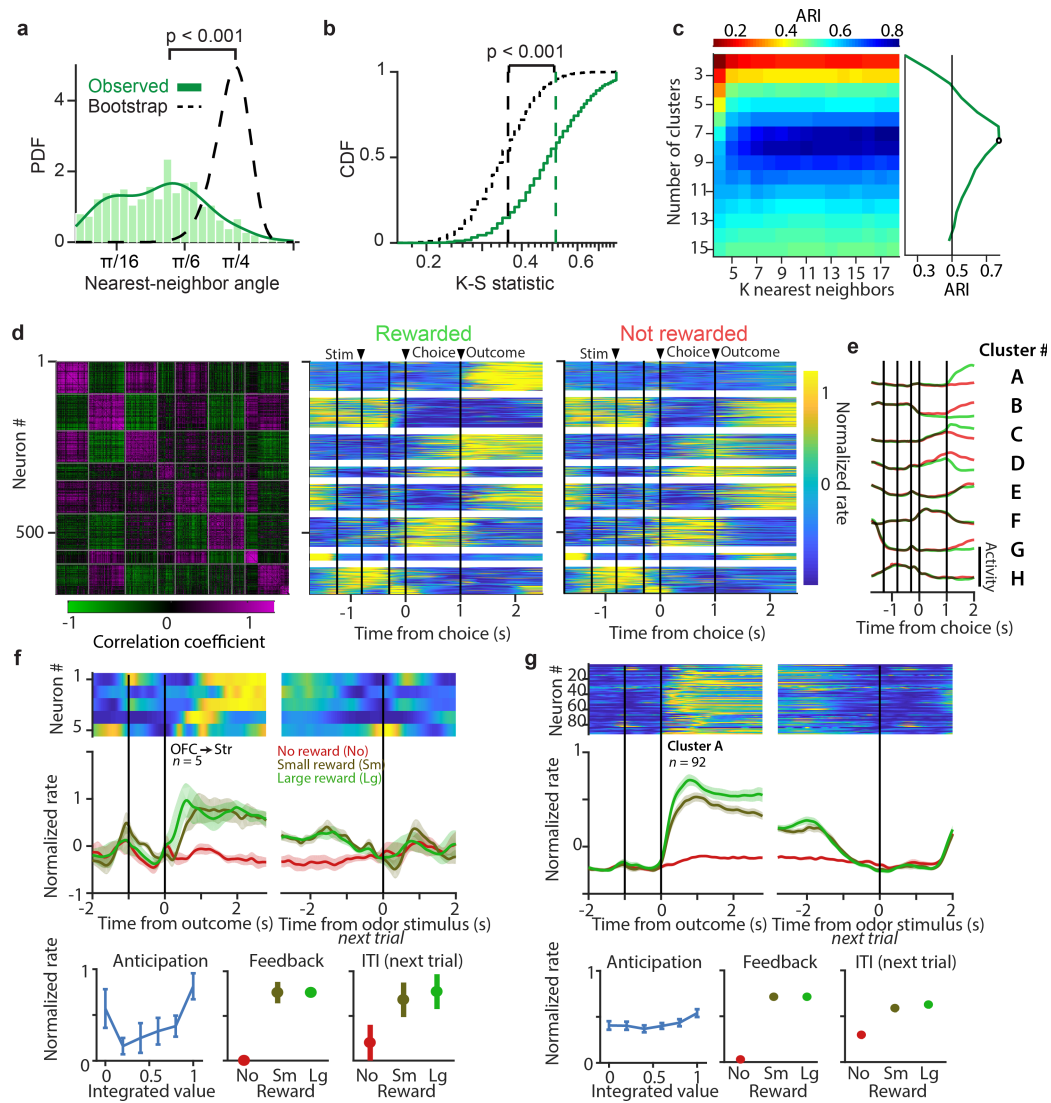
Extended Data Fig. 7 | Analysis of OFC response structure across behavioural epochs (cohort 1). **a**, Clustering results and hyperparameter selection for responses in the stimulus epoch (most stable configuration: $c = 5$ clusters and $k = 16$ nearest neighbours). Each combination of parameters is evaluated for stability using the adjusted rand index (ARI) (see Methods and Fig. 2). **b**, Clustering results for the stimulus epoch. The relationship between the five clusters can be examined visually by observation of the nearest-neighbour graph (top) and the within-cluster and between-cluster correlation coefficient (bottom). **c**, Clustering results and hyperparameter selection for responses in the feedback epoch (most stable configuration: $c = 5$ clusters and $k = 16$ nearest neighbours). Each combination of parameters is evaluated for stability using the adjusted rand index (ARI) (see Methods and Fig. 2). **d**, Clustering results for the feedback epoch. The relationship between the five clusters can be examined visually by observation of the nearest-neighbour graph (top) and the within-cluster and between-cluster correlation coefficient (bottom). **e–g**, Full cluster response profiles for all three epochs. **e**, Average response profiles of each of the 9 identified response clusters in the anticipation epoch (compare to Fig. 2). For each cluster, the normalized firing rate is shown for all 42 behavioural conditions used to generate the clustering results (responses conditioned on stimulus and choice, unbiased, left bias, and right bias blocks; conditioned on outcome of the previous choice and the evidence supporting the current choice, previous outcome). In addition, normalized firing rates are

shown conditioned on the size of the reward associated with the choice port (reward size). For each cluster, we also note the corresponding putative decision variable. **f**, Average response profiles of each of the five identified response clusters in the stimulus epoch. Conventions are the same as in **e**. Two of the clusters did not obviously map on a putative decision variable. **g**, Average response profiles of each of the five identified response clusters in the feedback epoch. Conventions are the same as in panel **e**. Two of the clusters did not map on a putative decision variable. **h**, **i**, Transition probabilities for neurons in a given cluster across subsequent epochs (compare to Fig. 5a). **h**, Transition probability for neurons belonging to a given cluster in the stimulus epoch to belong to a given cluster in the anticipation epoch (normalized per row). **i**, Transition probability between anticipation and feedback epochs (normalized per row). **j**, Neuron-based similarity measures across epochs. Neuronal response profiles are more similar across epochs for paired responses from the same neuron (red) compared to responses of two different neurons (black). Left, comparison of stimulus and anticipation epochs; right, comparison of anticipation and feedback epochs. Two-sample Kolmogorov–Smirnov test. **k**, Cluster-based similarity measures across epochs. Clusters derived from different epochs are more likely to share members if the average response profiles of each cluster are similar. Left, comparison of stimulus and anticipation epochs; right, comparison of anticipation and feedback epochs.



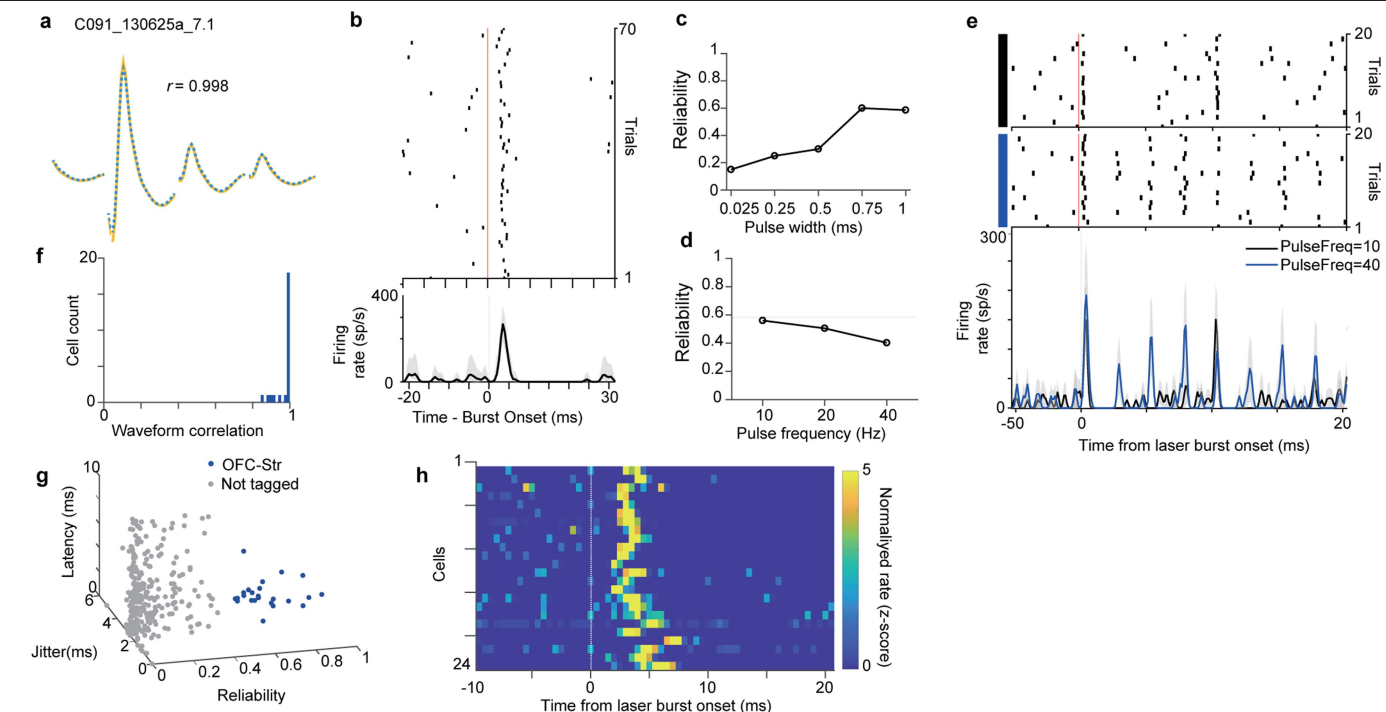
Extended Data Fig. 8 | Negative confidence is quantitatively represented in a cluster of OFC neurons. **a**, Response profile of neurons in cluster 2 correspond to the decision-variable confidence⁽⁺⁾. Panel (i): peri-stimulus time histogram of normalized firing rate, grouped based on the degree of evidence supporting choice. Panel (ii): trial-by-trial fit of each neuron in cluster 2 to choice evidence reveals significant representation of statistical decision confidence variable (R^2 with $P < 0.05$ based on bootstrap). Panels (iii–vi): normalized average tuning curves for neurons in cluster 2. Panel (iii) shows firing rate as a function of stimulus difficulty and choice (vevaimetric curve). Panel (iv) shows choice accuracy as a function of stimulus difficulty and firing rate (conditioned psychometric curve); (v) shows choice accuracy as a function of firing rate (calibration curve); and (vi) shows firing rate as a function of evidence supporting choice. Note that panels (iii) and (v) are replications of Fig. 4a bottom panels. **b**, Response profile of neurons in cluster 1, corresponding to a decision variable representation of confidence⁽⁻⁾ (same convention as panel **a**). **c**, Response profile of neurons in cluster 3, corresponding to a decision-variable of integrated value⁽⁻⁾. Panels (i–v): the representation of integrated value is analysed similarly to confidence in panel **a**, with the following changes: (ii): trial-by-trial fit of integrated value reveals significant representation of negative integrated value (R^2 with $P < 0.05$ based on bootstrap). (vi): firing rate

as a function of negative integrated value. (v): Choice accuracy as a function of firing rate ($*P < 0.01$, t -test). (vi): Firing rate as a function of integrated value. Note that panel (iii) and (v) are replications of Fig. 4b bottom panels. **d**, Single neurons encode coherent combinations of confidence and reward size. Each neuron's response profile was fit to a two-parameter model representing confidence and reward size. For most neurons, regression coefficients (β) for each component share the same sign. Data are shown for all neurons (grey), and neurons with significant beta coefficients for both components are shown in blue ($P < 0.01$ threshold). Polar histogram is significantly different from uniform ($P < 0.01$). **e**, Elementary task variables defined for the regression model. Each task variable was z-scored according to the weight of its non-masked conditions, with masked conditions subsequently set to zero. **f**, Detailed results of the LASSO model from Fig. 4g for neuronal clusters, using both the canonical design matrix (corresponding to decision variables) and null models (corresponding to random rotations of the design matrix). Errors are shown as median \pm s.e.m. P value calculated as paired sign-rank test. **g**, Detailed results of the LASSO model shown in Fig. 4h for single neurons for both the canonical design matrix (corresponding to decision variables) and null models (corresponding to random rotations of the design matrix). Errors are shown as median \pm s.e.m. P value calculated as a two-sided t -test.



Extended Data Fig. 9 | Time course clustering and positively outcome selective OFC-striatum projection neurons. **a, b**, The ePAIRS (**a**) and eRP (**b**) tests reveals significant non-random clustering in the OFC population based on response profiles with temporal but not task-related information (see Methods). For ePAIRS, nearest neighbour angles were smaller than expected, suggestive of clustering (rather than dispersion). **c**, Clustering results and hyperparameter selection for temporal clustering (most stable configuration: $c = 8$ clusters and $k = 13$ nearest neighbours; compare to Fig. 3f; see Methods). **d**, Analysis of temporal response profiles (left); spectral clustering of temporal response profiles without tuning information reveals eight clusters with high within-cluster similarities ($n = 7$ rats combined; Methods). Dynamics of the trial-averaged time course for single neurons in the eight clusters for rewarded trials (middle) and error trials (right) are shown. We separated rewarded and error trials for this analysis as the actions performed

during the outcome period are very different (drinking water versus return to centre port). **e**, Average dynamics of the trial-averaged time course for the eight clusters (green, rewarded trials; red, error trials). **f**, Average activity in correct trials (top) and average PSTH grouped by outcome (error, small reward, large reward) of identified OFC-striatum projecting neurons that positively encoded outcome. Lower panels show that neurons are positively tuned to integrated value in the anticipation period and positive tuning to outcome in the feedback epoch and ITI. Conventions are the same as in Fig. 5h. **g**, Average PSTH of neurons in cluster A whose dynamics match those of optogenetically identified neurons encoding positive outcome (excluding optogenetically identified OFC-striatum projection neurons). Note that the coding of integrated value is weaker than for the negative population but still significant ($r(90) = 0.1, P = 0.02$). Conventions are the same as in panel **f**.



Extended Data Fig. 10 | Optogenetic identification of OFC-striatum

projection neurons. a–e, An example neuron showing reliable light-evoked responses. **a,** For an example neuron, average waveforms of spontaneous (yellow) and light-evoked spikes (blue) across four tetrodes are very similar. **b,** Spike raster (top) and PSTH (bottom) for the light-activated cell in panel **a** aligned to light onset (1 ms duration, first stimulus in a train). **c,** Reliability of the evoked responses to the first stimulus as a function of pulse duration. **d,** Probability of light-evoked spikes as a function of stimulation frequency (1 ms duration, 20 repetitions). **e,** Spike raster (top) and PSTH (bottom) aligned to

light onset for stimulation trials at 10 Hz and 40 Hz. **f,** Histogram of Pearson's correlation coefficients between the waveforms of spontaneous and light-evoked for identified OFC-striatum projecting neurons. **g,** Quantification of light-evoked responses, showing latency and jitter of light-evoked spikes for tagged neurons as a function of the reliability of evoking a response to light. Putative OFC-striatum projection neurons are shown (blue points). **h,** z-scored PSTH of all identified OFC-striatum projection neurons in response to 1–3 ms blue light stimulation.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Spike data was collected using NEURALYNX software, with spike sorting using MCLUST.

Data analysis

Code for ePAIRS and eRP analyses are available at <https://github.com/agvaughan/EllipticalClustering>.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Neuronal data from all experiments is available on reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes are described by number of neurons, and number of unique animals. Although there is no clear guideline for the number of animals or neurons required to validate a given result, we have attempted to maximize the reliability of our result by performing a full replicate of our analysis on a second cohort of animals.
Data exclusions	No animals were excluded from analysis. Neurons were selected according to strict spike-sorting criteria, as described in the manuscript; we note that any failures of spike sorting (i.e. multi-unit recordings) would generally bias our results against the conclusions we observe.
Replication	As discussed in the manuscript, we replicated our main result through a complete re-analysis of data drawn from a new cohort of animals trained separately.
Randomization	Not applicable as there is no group allocation in this study.
Blinding	Not applicable as there is no group allocation in this study. However, criteria for analysis of replicate data were determined before analysis (and matched the original analysis as closely as possible).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Male Long Evans rats (~250 g) were used for all experiments.
Wild animals	N/A
Field-collected samples	N/A
Ethics oversight	All procedures involving animals were approved by the Cold Spring Harbor Laboratory Institutional Animal Care and Use Committee and by the Animal Research Committee of Doshisha University, and were carried out in accordance with National Institutes of Health standards.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Chimeric peptidomimetic antibiotics against Gram-negative bacteria

<https://doi.org/10.1038/s41586-019-1665-6>

Received: 8 August 2018

Accepted: 13 August 2019

Published online: 23 October 2019

There are amendments to this paper

Anatol Luther^{1,6}, Matthias Urfer^{2,6}, Michael Zahn³, Maik Müller⁴, Shuang-Yan Wang², Milon Mondal², Alessandra Vitale⁵, Jean-Baptiste Hartmann³, Timothy Sharpe³, Fabio Lo Monte², Harsha Kocherla², Elizabeth Cline², Gabriella Pessi⁵, Parthasarathi Rath³, Seyed Majed Modaresi³, Petra Chiquet¹, Sarah Stiegeler¹, Carolin Verbree¹, Tobias Remus¹, Michel Schmitt¹, Caroline Kolopp¹, Marie-Anne Westwood¹, Nicolas Desjonquères¹, Emile Brabet¹, Sophie Hell¹, Karen LePoupon¹, Annie Vermeulen¹, Régis Jaisson¹, Virginie Rithié¹, Grégory Upert¹, Alexander Lederer¹, Peter Zbinden¹, Achim Wach¹, Kerstin Moehle², Katja Zerbe², Hans H. Locher¹, Francesca Bernardini¹, Glenn E. Dale¹, Leo Eberl⁵, Bernd Wollscheid⁴, Sebastian Hiller³, John A. Robinson^{2*} & Daniel Obrecht^{1*}

There is an urgent need for new antibiotics against Gram-negative pathogens that are resistant to carbapenem and third-generation cephalosporins, against which antibiotics of last resort have lost most of their efficacy. Here we describe a class of synthetic antibiotics inspired by scaffolds derived from natural products. These chimeric antibiotics contain a β -hairpin peptide macrocycle linked to the macrocycle found in the polymyxin and colistin family of natural products. They are bactericidal and have a mechanism of action that involves binding to both lipopolysaccharide and the main component (BamA) of the β -barrel folding complex (BAM) that is required for the folding and insertion of β -barrel proteins into the outer membrane of Gram-negative bacteria. Extensively optimized derivatives show potent activity against multidrug-resistant pathogens, including all of the Gram-negative members of the ESKAPE pathogens¹. These derivatives also show favourable drug properties and overcome colistin resistance, both in vitro and in vivo. The lead candidate is currently in preclinical toxicology studies that—if successful—will allow progress into clinical studies that have the potential to address life-threatening infections by the Gram-negative pathogens, and thus to resolve a considerable unmet medical need.

The rapid emergence of antimicrobial resistance is now a matter of global concern, and public awareness of an emerging crisis has been highlighted in several recent reports^{1–3}. According to the World Health Organization (WHO), *A. baumannii*, *P. aeruginosa* and Enterobacteriaceae (which are Gram-negative members of the so-called ESKAPE (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter* spp.) pathogens¹) that are resistant to carbapenem or third-generation cephalosporins are of particular concern. Novel antibiotics against Gram-negative bacteria are urgently needed, particularly because resistance against colistin—an antibiotic of last resort—is on the rise globally⁴.

The outer membrane of Gram-negative bacteria comprises an asymmetric bilayer, with glycerophospholipids in the inner leaflet and lipopolysaccharide (LPS) in the outer leaflet⁵. This unique permeability barrier protects the bacteria from toxic environmental factors (such as antibiotics) and contains many integral β -barrel outer membrane proteins (OMPs), which are required for biogenesis of the outer membrane⁶.

One family of new macrocyclic β -hairpin peptidomimetic antibiotics has previously been reported that target the β -barrel OMP LptD, specifically in *Pseudomonas* spp.^{7,8}. Murepavadin (formally, POL7080) (Fig. 1) is the first OMP-targeting antibiotic to enter late-stage clinical development.

Here we report the discovery of a family of chimeric peptidomimetic antibiotics that possess broad-spectrum antimicrobial activity against Gram-negative bacteria. We provide evidence that these antibiotics have a mechanism of action that involves binding to LPS and the essential OMP BamA, which is the main component of the β -barrel folding complex (BAM) that promotes folding and insertion of β -barrel proteins into the outer membrane in Gram-negative bacteria^{6,9}.

Discovery and antibacterial activity

Screening cyclic peptides related to murepavadin (1, Fig. 1) on a panel of Gram-negative ESKAPE pathogens resulted in initial hits (including

¹Polyphor AG, Allschwil, Switzerland. ²Chemistry Department, University of Zurich, Zurich, Switzerland. ³Biozentrum, University of Basel, Basel, Switzerland. ⁴Institute of Molecular Systems Biology & Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland. ⁵Department of Plant and Microbial Biology, University of Zurich, Zurich, Switzerland. ⁶These authors contributed equally: Anatol Luther, Matthias Urfer. *e-mail: john.robinson@chem.uzh.ch; daniel.obrecht@polyphor.com

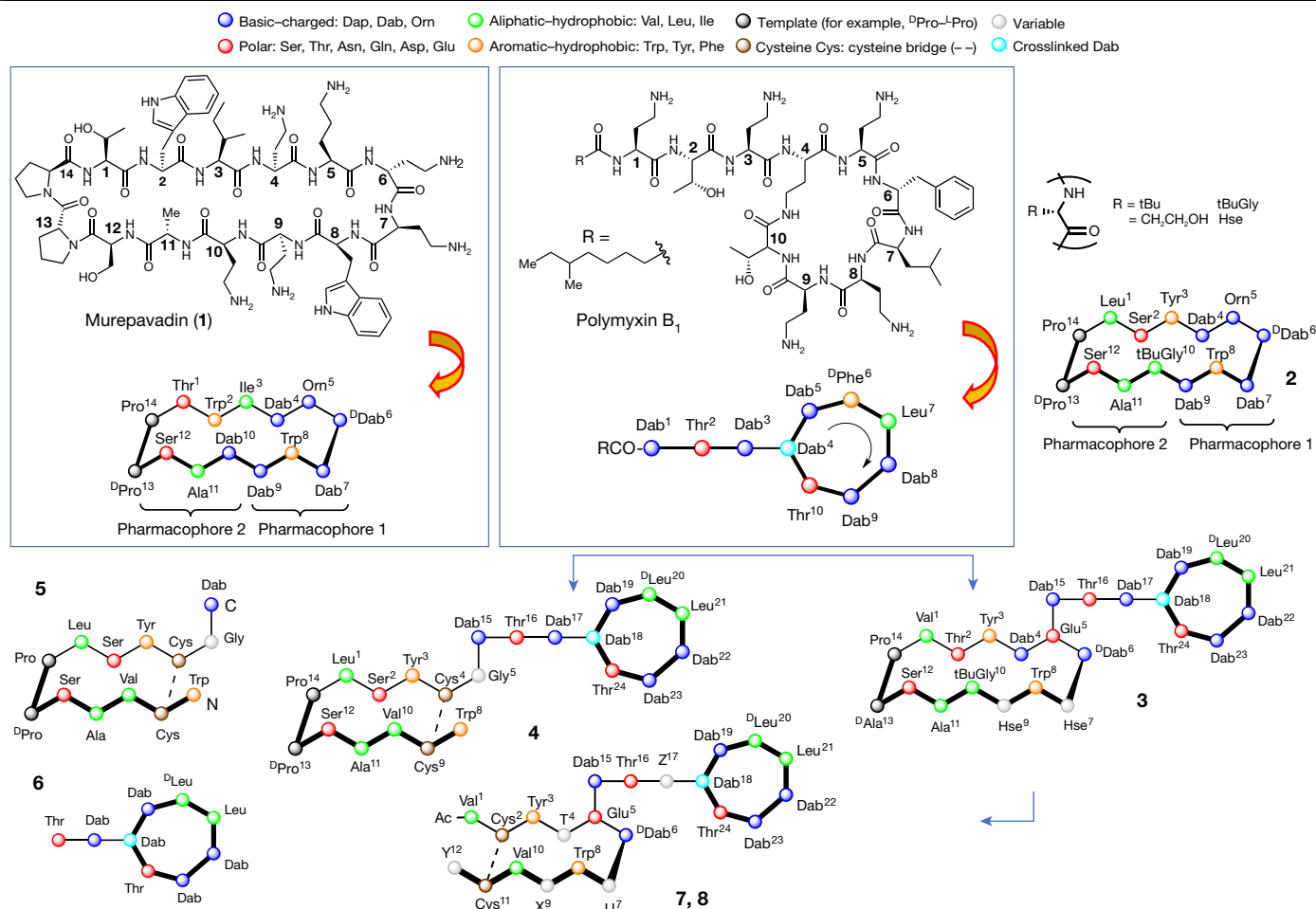


Fig. 1 | Structures of the chimeric antibiotics. The structures of murepavadin and polymyxin B₁ are shown in the boxes, along with cartoons that represent these structures. For structures of all of the peptides used in this work, see

Extended Data Fig. 10 and Extended Data Table 1. The corresponding cartoons are also used to represent the structures of peptides 2–8.

peptide **2**) with promising antimicrobial activities (range of minimal inhibitory concentrations (MICs) of 2–8 mg l⁻¹) (Table 1). These hits also showed activity against colistin-resistant strains, but not against Gram-positive bacteria (*S. aureus*) and fungi (*Candida* spp.). However, **2** exhibited substantially higher MICs in the presence of 50% human serum (Supplementary Table 2, Supplementary Information), and residual membrane lytic activity against human red blood cells was also observed (Supplementary Table 3). We therefore synthesized a series of chimeric molecules (see Supplementary Information for experimental details), in which the β -hairpin macrocycle was linked to the peptide macrocycle of polymyxin B (PMB) and colistin¹⁰. Because the peptide macrocycles in PMB and colistin bind to the lipid A portion of LPS¹¹, we hypothesized that this property might synergize with the OMP-targeting activity of a β -hairpin motif, if both components were combined into chimeric molecules.

Promising results were obtained with chimaeras **3** and **4** (Fig. 1; residue numbering is the same as is used in **1** and **2**, to aid with comparisons), which showed potent activity against a panel of Gram-negative ESKAPE pathogens (Table 1, Supplementary Table 2). However, peptide **5** and PMB nonapeptide (**6**) (Fig. 1), which represent the individual parts of peptide **4**, showed no antimicrobial activity (MICs > 32 mg l⁻¹). The enantiomer of **4** (**e-4**) showed about 30-fold lower activity, suggesting a chiral target (Supplementary Table 2). Compound **4** was not investigated further, owing to relatively high protein binding (Supplementary Table 3) and unfavourable pharmacokinetic and pharmacodynamic

properties in vivo. By contrast, compound **3** showed favourable in vitro and in vivo properties (Table 1, Extended Data Fig. 1, Supplementary Table 2), maintained its activity in the presence of 50% human serum (MIC shift ≤ 4 -fold) (Supplementary Table 2), and showed excellent in vivo efficacy in several mouse models of infection (Extended Data Fig. 1). Compound **3** was active against the wild-type *A. baumannii* strain as well as an isogenic LPS-deficient strain ($\Delta lpxA$) (Supplementary Table 2), consistent with an interaction with both LPS and an alternative interaction target on these cells.

A substantial medicinal-chemistry effort resulted in the synthesis of a family of compounds in which the hairpin structure is stabilized by a disulfide bond, as represented by **7** and **8** (Fig. 1). Both **7** and **8** show excellent in vitro activity against Gram-negative ESKAPE pathogens, including multidrug-resistant, extensively drug resistant and colistin-resistant isolates, with a MIC at which 90% of the isolates were inhibited (MIC₉₀) in the range 0.06–0.25 mg l⁻¹ (Table 1, Supplementary Table 2). A control, peptide **9** (Extended Data Table 1), with a scrambled sequence was inactive (MIC ≥ 64 mg l⁻¹). The activity of enantiomer **e-8** was even lower than that of **e-4** (Supplementary Table 2). No activity was observed against Gram-positive *S. aureus* or against strains that are intrinsically resistant to antimicrobial peptides, such as *Proteus* spp. and *Serratia marcescens*.

Peptides **3** and **8** showed a rapid bactericidal activity, with more than three log₁₀ reductions in colony-forming units (CFUs) observed within 2 h at 1–4 \times MIC (Extended Data Fig. 1). The potential for resistance

Table 1 | Antimicrobial activities

	2	3	4	7	8	Meropenem	Ceftazidime	Tobramycin	Colistin
<i>A. baumannii</i> A369	1	0.06	0.06	0.06	0.06	>64	>64	8	0.25
<i>A. baumannii</i> 863866	4	0.25	0.06	0.25	0.06	32	>64	4	64
<i>A. baumannii</i> 872842	4	0.13	0.06	0.25	0.06	>8	>8	0.25	>8
<i>P. aeruginosa</i> UU6419	8	0.5	0.5	0.25	0.25	64	>64	>64	0.5
<i>P. aeruginosa</i> 22409	>8	0.5	2	0.5	0.25	32	>64	8	1
<i>P. aeruginosa</i> 401190	2	0.25	0.13	0.13	0.13	>64	>64	>64	0.5
<i>E. cloacae</i> 867213	8	0.13	0.13	0.25	0.13	≤0.06	>64	16	>64
<i>E. cloacae</i> 950265	1	0.13	0.06	0.13	0.06	0.13	64	>64	8
<i>E. coli</i> 959670	1	0.25	0.25	0.25	0.06	≤0.06	64	32	4
<i>E. coli</i> 402788	0.5	0.06	0.06	0.06	0.03	64	>64	>64	0.13
<i>E. coli</i> 926415	0.5	0.13	0.13	0.25	0.13	0.03	>8	>8	8
<i>K. pneumoniae</i> 403575	2	0.13	0.13	0.25	0.13	64	>64	16	4
<i>K. pneumoniae</i> 946897	4	0.5	0.5	2	0.25	>64	>64	16	16
<i>K. pneumoniae</i> RV 9959	1	0.13	0.06	0.13	0.06	32	>64	16	1
<i>S. aureus</i> ATCC 29213	64	>64	>64	>64	>64	0.13	>8	>8	>8

The MICs (in mg l⁻¹) of the chimeric antibiotics against the strains indicated were measured by the Clinical and Laboratory Standards Institute (CLSI) microbroth method, along with those of a panel of commercial antibiotics as a comparison. Values in bold denote sensitivity; values without bold denote resistance (Methods). No antimicrobial activity for the chimeric antibiotics (MICs ≥ 64 mg l⁻¹) was seen against *S. aureus*.

development against **8** was assessed by plating 10⁹–10¹⁰ CFUs of different strains on Mueller–Hinton (MH) agar with antibiotic concentrations up to 64 mg l⁻¹. As no mutants with increased MICs could be obtained, serial passage experiments in liquid medium were performed with increasing concentrations of **8**. In this way, some resistant mutants of *K. pneumoniae* SSI3010 with an increased MIC against **8** were isolated (Extended Data Fig. 1c).

Biological profile

Peptides **3**, **7** and **8** displayed low toxicity toward mammalian cells (HeLa cells) (Supplementary Table 3a). No general membrane lytic activity was observed towards red blood cells, and all three compounds showed favourable plasma protein binding and human plasma stability (Supplementary Table 3a). The in vivo tolerability and pharmacokinetics profile in mice of peptides **7** and **8** were also favourable (Supplementary Table 3b).

Compounds **3**, **7** and **8** were evaluated in various neutropenic mouse models of infection, including septicaemia, peritonitis and thigh infections of *A. baumannii*, *Escherichia coli*, *K. pneumoniae* and *P. aeruginosa*. Compound **3** showed potent in vivo efficacy, in a mouse model of septicaemia, against the *E. coli* 5799 clinical isolate (Extended Data Fig. 1d). The in vivo efficacy of **3** and **8** was demonstrated, in neutropenic mouse models of peritonitis, against *K. pneumoniae* SSI3010 and *E. coli* SNTR36B6, which contains *mcr-3* (Extended Data Fig. 1g, h). A good efficacy of **8** was shown in a mouse model of peritonitis with the colistin-resistant *E. coli* AF45 isolate, which contains *mcr-1* (Extended Data Fig. 1e, f). The dose response of **8** was investigated in neutropenic mouse models of thigh infection, in which the potent efficacy of **8** was confirmed against *P. aeruginosa* PA14 and the extensively drug-resistant *A. baumannii* NCTC 13301 (Extended Data Fig. 1j, k). Similar results were observed for *E. coli* ATCC BAA 2469, an NDM-1 carbapenem-resistant isolate (Extended Data Fig. 1i). The propensity to generate nephrotoxicity was assessed in vivo in an acute mouse model¹². Whereas **7** and **8** showed zero-to-mild nephrotoxicity (scores of 1 and 2, respectively), colistin—as expected—showed a higher nephrotoxicity, with a score of 24 (Supplementary Table 3c).

Mechanism-of-action studies

No effects of **3** and **4** were detected on protein, RNA, DNA or cell-wall biosynthesis in *E. coli* ATCC 25922 (Extended Data Fig. 2). However, **3** and **4** permeabilized the *E. coli* cell envelope, as shown using the nucleic-acid stain SYTOX-Green⁷ in cells treated with **3** or **4** (Extended Data Fig. 3a). These permeabilizing effects of **3** and **4** were also seen for the clinical *E. coli* strain 926415, which is resistant to PMB and colistin (MIC for PMB > 32 mg l⁻¹)—although PMB lost its cell-envelope permeabilizing effect on this strain (Extended Data Fig. 3b).

The chimaeras **3** and **4** caused the rapid release of periplasmic β-lactamase and cytoplasmic β-galactosidase (Extended Data Fig. 3c, d), which confirms that **3** and **4** permeabilize both the inner and outer membrane in *E. coli* (including in PMB-resistant strains). The major forms of lipid A in these PMB-resistant strains were analysed, which confirmed the presence of lipid A modified with phosphoethanolamine and/or 4-amino-4-deoxyarabinose (Ara4N)^{13,14} (Extended Data Fig. 4). Thus, these lipid A modifications¹⁵ are not sufficient to confer resistance against the chimeric compounds. We also examined inducible PMB resistance in *E. coli* ATCC 25922 grown in low Mg²⁺ concentrations¹⁶, which activates PhoP–PhoQ¹⁷. The MICs for **3** and **4** were also unaffected by growth in low Mg²⁺ (Supplementary Table 2e).

Super-resolution stimulated emission depletion (STED) fluorescence microscopy was used with *E. coli* ATCC 25922 that was exposed to the antibiotic and the membrane dye FM4-64, the nucleic-acid stain DAPI and SYTOX-Green to detect permeabilized cells. The nucleoids stained with DAPI were not influenced substantially by **3** and **4** (Fig. 2a–c), although many cells treated with **3** or **4** showed strong green fluorescence in the presence of SYTOX-Green, which again reveals the permeabilizing effects on the inner and outer membrane (Fig. 2a–c). Some of the treated cells showed membrane-associated red fluorescence puncta upon staining with FM4-64, suggesting perturbations to membrane structure.

The coupling of **3** and **4** to Alexa Fluor 488 produced the derivatives **fl-3** and **fl-4** (Extended Data Table 1), which showed good antimicrobial activity (MICs versus *E. coli* ATCC 25922 of 0.1 mg l⁻¹ for **fl-3**, and 2.0 mg l⁻¹ for **fl-4**). STED fluorescent imaging with **fl-3**, **fl-4** and *E. coli* ATCC 25922 showed the accumulation of fluorescence into bright green spots in the membrane, which suggests a predominant site of

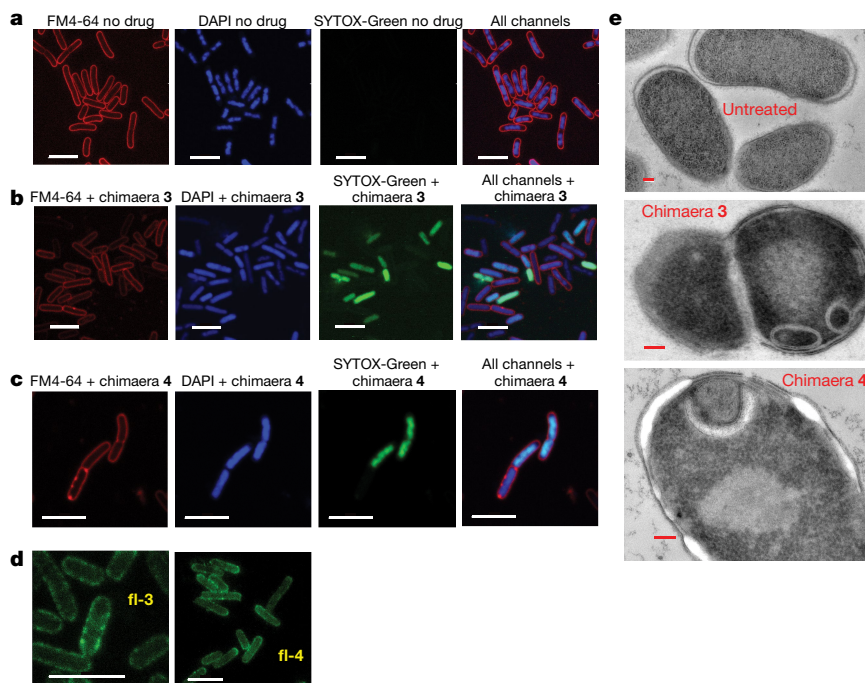


Fig. 2 | Fluorescence and electron microscopy. **a–d**, Fluorescence microscopy of *E. coli* ATCC 25922 cells grown in MH-II and stained with FM4-64, DAPI or SYTOX-Green, without treatment (**a**) or grown with chimaera **3** (0.3 mg l^{-1}) (**b**), or with chimaera **4** (0.6 mg l^{-1}) (**c**). Cells grown without drug in MH-II and stained with the fluorescently labelled antibiotics (**fl-3** or **fl-4**) are shown in **d**. A Leica CLSM SP8 gSTED microscope was used. Scale bars, $4 \mu\text{m}$. **a** and **c** show a

representative example of $n = 3$ biologically independent experiments; **b** and **d** show a representative example of $n = 2$ biologically independent experiments. **e**, Transmission electron microscopy of *E. coli* ATCC 25922 untreated or grown with **3** or **4** at concentrations causing about 50% growth inhibition (about 0.1 mg l^{-1}) ($n = 3$ biologically independent experiments). Scale bars, 200 nm .

interaction of the chimaeras with OMP clusters in the cell envelope (Fig. 2d).

E. coli treated with **3** or **4** led to marked changes in membrane structure that were revealed by transmission electron microscopy, which showed the presence of extra membrane-like material, membrane detachment and the appearance of vacuoles (Fig. 2e). For the treatment with **4**, the appearance of bright areas in the cytoplasm were also noted. Using scanning electron microscopy, cells treated with **3** and **4** showed collapsed membranes and extracellular knob-like structures (Extended Data Fig. 5), which suggests that both of the chimaeras perturb bacterial membranes.

Photo-affinity interaction mapping was used to search for interaction partners for **3**, **4** and **7** in the cell membrane. The photoprobes **PAL-3**, **PAL-4** and **PAL-7** (Extended Data Table 1) retain good antimicrobial activity against *E. coli* ATCC 25922 (**PAL-3**, MIC of about 0.1 mg l^{-1} ; **PAL-4**, MIC of about 1 mg l^{-1} ; and **PAL-7**, MIC of about 0.1 mg l^{-1}). For the treatment with **PAL-4**, photolabelling of whole cells revealed the labelling of proteins of about 90 kDa mass, as well as multiple components in the range of 25 to 50 kDa . Similar results were obtained with **PAL-3** and **PAL-7** (Extended Data Fig. 6a).

For the identification of **PAL-3**, **PAL-4** and **PAL-7** binding proteins, we combined photo-crosslinking with target affinity purification, proteolytic digestion and mass-spectrometry-based proteomics, using cells that were treated in the same way with **3**, **4**, or **7** as controls. In total, $1,320$ proteins from *E. coli* were label-free-quantified with at least two peptides at a false discovery rate below 1% . Relative quantitative comparisons revealed the specific and photolabelling-dependent enrichment of several proteins with a subcellular localization at the outer membrane, on the basis of UniProt annotations (Fig. 3). For the treatment with **PAL-4**, the OMPs BamA and LptE were significantly enriched, whereas for **PAL-3** treatment BamA, BamD and LamB were

significantly enriched. Similarly, **PAL-7** labelling studies confirmed that BamA, BamD, LptE and LamB were significantly enriched OMPs (Extended Data Fig. 6b); BamA was the only outer membrane protein that was consistently labelled by all three photoprobes. BamA (about 90 kDa) and BamD (about 28 kDa) are both essential components of the BAM foldase complex⁶. LptE (about 21 kDa) is an essential component of the LptD–LptE complex that is required for LPS transport to the cell surface¹⁸, whereas LamB (about 50 kDa) functions as an outer-membrane maltose transporter. The highly abundant OmpA and OmpF porins in *E. coli* were not significantly enriched in these experiments.

The interaction between **3** and BamA was examined using in vitro binding experiments with Cy3-labelled **3** (**Cy3-3**) (Extended Data Table 1), the BamA N-terminal periplasmic POTRA domains and the BamA C-terminal β -barrel domain, each as a recombinant protein¹⁹. Although the peptide showed no detectable interaction with BamA(POTRA1–5) (Extended Data Fig. 7a, c), both fluorescence anisotropy and microscale thermophoresis showed the binding of the antibiotic to the β -barrel domain of BamA (BamA- β) with a dissociation constant (K_D) in the sub-micromolar range (Fig. 4c, Extended Data Fig. 7c). No binding was seen to lauryldimethylamine oxide (LDAO) micelles (Extended Data Fig. 7a), and no binding of **Cy3-3** was observed with LamB and OmpX (Extended Data Fig. 7b). Additional fluorescence-anisotropy binding experiments were performed between a Cy3-labelled control peptide **9** (Extended Data Table 1) and BamA- β , but again no binding was detected (Extended Data Fig. 7b).

The interaction site of peptide **3** with BamA was mapped by high-resolution nuclear magnetic resonance (NMR) spectroscopy. BamA in LDAO detergent and lipid bilayer nanodiscs populates a dynamic ensemble of open and closed states that can be detected by NMR

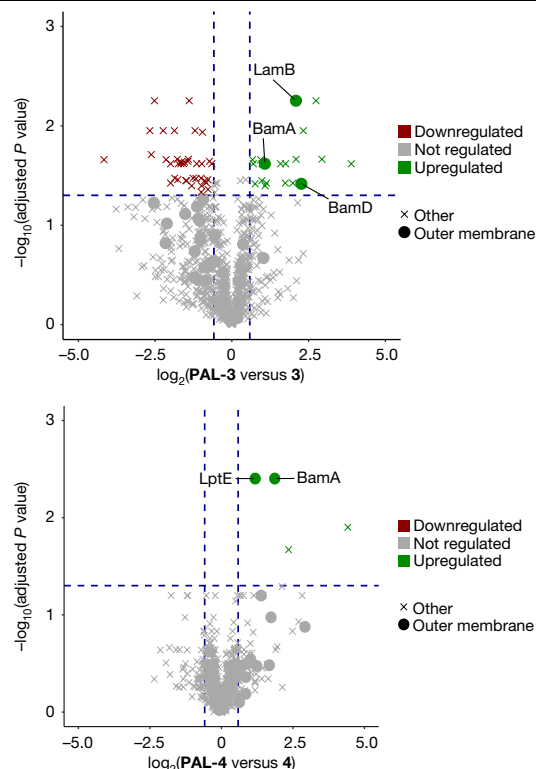


Fig. 3 | Photo-affinity interaction mapping. Volcano plots showing the relative abundance of proteins that were streptavidin-captured from *E. coli* cells photolabelled with **PAL-3** or **PAL-4** ($n = 3$ biologically independent samples each) versus control cells treated with **3** or **4** ($n = 4$ biologically independent samples each). Fold changes in protein abundance (expressed in \log_2) were calculated by linear mixed-effect model, and tested for statistical significance using a two-sided t -test. The P values that were obtained were further corrected for multiple comparisons using the Benjamini–Hochberg method. Proteins are represented based on UniProt-annotated subcellular location as dots (outer membrane) or crosses (no, or other, location). Significantly enriched proteins (with an abundance ratio ≥ 1.5 and adjusted $P \leq 0.05$; these thresholds are shown as blue lines) are highlighted in green, and represent candidates labelled with **PAL-3** (top) or **PAL-4** (bottom). A full list of proteins quantified by mass spectrometry in these experiments is supplied as Source Data.

spectroscopy¹⁹. The binding site of peptide **3** was determined by chemical shift perturbations in two-dimensional [^{15}N , ^1H]transverse relaxation-optimized spectroscopy (TROSY) spectra with BamA^{ext} (a variant of the BamA barrel that is locked in the closed state¹⁹). Significant ligand-induced chemical shift perturbations for selected resonances were observed in the extracellular loops L4, L6 and L7 of BamA^{ext}, but none of these perturbations was observed within the β -barrel domain (Fig. 4, Extended Data Fig. 8). A titration of peptide **3** with non-stabilized BamA- β showed that the interaction of the peptide shifted the conformational ensemble, stabilizing the closed form of BamA (Fig. 4). Analogous experiments with **8** and the inactive, scrambled peptide **9** revealed similar chemical shift perturbations for **8**, which were localized mainly in the extracellular loops (L4, L6 and L7) of BamA^{ext}; these experiments also confirmed the ability of compound **8** to stabilize the closed state of BamA (Extended Data Figs. 8, 9). By contrast, the microbiologically inactive scrambled peptide **9** led to only weak chemical shift perturbations (Extended Data Figs. 8c, 9d). In particular, several residues that showed strong chemical shift perturbations with the active peptide (Fig. 4, Extended Data Fig. 8a, b) showed no chemical shift perturbations with the scrambled peptide **9**.

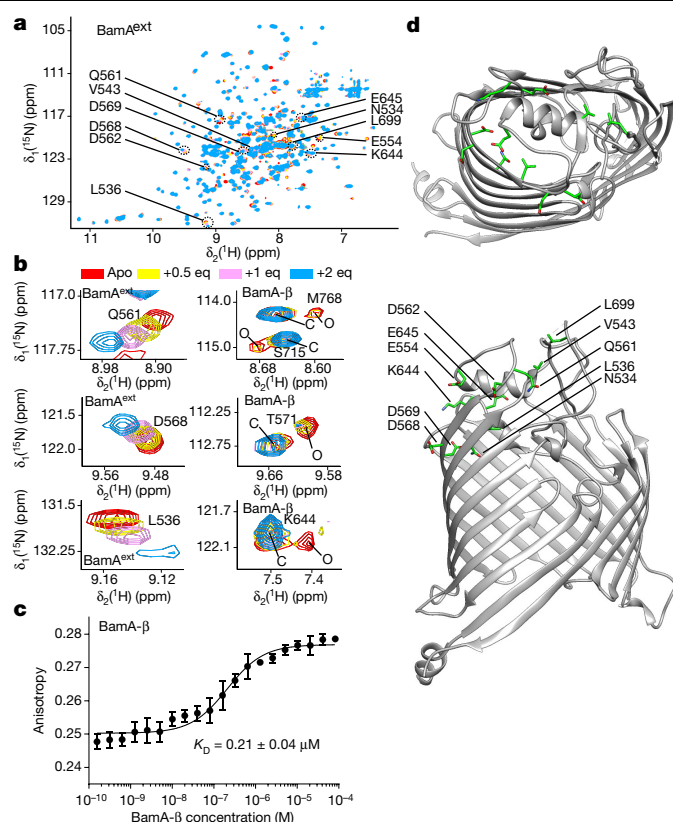


Fig. 4 | Antibiotic binding to the BamA β -barrel. The interaction of **3** with the BamA β -barrel domain was characterized using NMR spectroscopy and fluorescence anisotropy. **a**, Two-dimensional [^{15}N , ^1H]TROSY spectra of 300 μM [^{15}N , ^{15}N]BamA^{ext} (red) titrated with 0.5 (yellow), 1 (magenta) or 2 (blue) equivalents (eq) of peptide **3**. Measurements were performed once. **b**, Close-up views of selected residues from the titration in **a** and from a corresponding titration with BamA- β . C, closed conformation; O, open conformation. **c**, Fluorescence anisotropy measurement of the binding of peptide **Cy3-3** to BamA- β . Measurements were performed in triplicate. Error bars show standard deviations around the mean. **d**, Representation of the interactions of peptide **3** on the BamA β -barrel structure viewed from the top and from the side of the barrel. Labelled residues (in green) have substantial chemical shift perturbations or intensity changes upon peptide binding (crystal structure from Protein Data Bank code (PDB) 6FSU).

Resistant mutants were isolated after multiple passages of *K. pneumoniae* SSI3010 with peptide **8** (Extended Data Fig. 1c). Whole-genome sequencing of mutants with increased MICs (Supplementary Table 4) revealed that mutants from early rounds with increases in MIC that were only modest, contained mutations in genes involved in the regulation (*phoQ*) and formation of modified lipid A (*arnC*). Mutations in these genes lead to high levels of resistance against colistin (MIC = 64 mg l^{-1}); however, here they led to only a modest MIC increase against **8** (from 0.06 to 1 mg l^{-1}) (Supplementary Table 4). When resistant mutants were selected with colistin, single mutations in *phoQ* were sufficient to confer high resistance to colistin (MICs increasing from 0.25 to 8 or 64 mg l^{-1}) but had only a small effect on the MICs of **8** (increasing from 0.06 to 0.125 or 0.5 mg l^{-1}) (Supplementary Table 4). In the final passages with **8**, two different isolates of *K. pneumoniae* SSI3010 were obtained that showed a large increase in MIC (128 mg l^{-1}), and which had different colony morphologies on agar plates. These isolates contained (among others) a mutation in *bamA* that corresponds to a D703Y exchange in the external loop L6 of the BamA β -barrel (Supplementary Table 4). Complementation experiments were conducted with genes that encode wild-type BamA (*bamA*) and the BamA(D703Y) mutant (*bamA*^{D703Y}). Isolates from passages 8 and

13—as well as the parental strain that expresses wild-type *bamA*—were transformed with a plasmid that expresses BamA(D703Y), whereas an isolate from passage 16 that produces BamA(D703Y) was transformed with a plasmid that expresses wild-type BamA. The complementation with wild-type BamA in the passage-16 isolate led to a fourfold decrease in MIC, which indicates that BamA indeed has a role in the antimicrobial activity of compound **8**. Complementation with BamA(D703Y) in the isolates from passages 8 and 13 led to a significant ($P > 0.01$) increase in the MIC (Supplementary Table 5). However, complementation of the parental strain with BamA(D703Y) did not lead to a significant ($P > 0.01$) increase in the MIC, possibly owing to a dominant effect of the wild-type gene present in this bacterium. Overall, the genetic studies complement our binding studies, and provide support for the involvement of BamA as a binding target for compound **8**.

Discussion

The compounds **3**, **4** and **7** (and the closely related analogue, **8**) are bactericidal against a broad panel of Gram-negative ESKAPE pathogens (including multi-drug resistant, extensively drug resistant and colistin-resistant strains), show low cytotoxicity towards mammalian cells, have a low propensity to elicit resistance in all of the bacterial strains that we tested, maintain a high potency in the presence of human serum, and show favourable safety and pharmacokinetic properties. This translates into potent in vivo efficacy in various mouse models of peritonitis, including infections induced with colistin-resistant *E. coli* strains that contain *mcr-1* and *mcr-3* resistance genes and mouse models of thigh infection with *E. coli* ATCC BAA 2469 (an NDM-1 strain), *A. baumannii* NTCT 13301 (an extensively drug resistant strain) and *P. aeruginosa* PA14. The low propensity of **7** and **8** to generate renal toxicity, in a comparison with colistin, is of particular note.

Exposure of *E. coli* cells to the chimaeras leads to permeabilization of both the inner and outer membrane (Fig. 2, Extended Data Fig. 3) and causes substantial perturbations to membrane architecture (Fig. 2, Extended Data Fig. 5). A key question is how the permeabilizing effects arise, given that the compounds show no general lytic activity on typical eukaryotic membrane bilayers that are composed of glycerophospholipids.

One interaction target identified by photo-affinity interaction analyses for all of the tested photoprobes was BamA, the main component of the BAM complex in *E. coli*⁶. Other proteins identified by individual photoprobes include β -barrel OMPs (for example, LamB and LptD) or lipoproteins that are closely involved in β -barrel OMP folding and outer membrane biogenesis (for example, BamD and LptE). Indeed, both LamB and LptD–LptE are heavily dependent on the BAM complex for folding and insertion into the outer membrane^{20,21}. The appearance of multiple OMPs in the pull-down experiments could be the result of diffusion of the photo-activated probe within OMP clusters that are known to predominate in the outer membrane²².

The identification of BamA as a binding target for the chimaeras was confirmed conclusively by in vitro studies, in which **3** was shown to bind with high selectivity to BamA- β (Fig. 4, Extended Data Fig. 7). Furthermore, chemical shift mapping by NMR spectroscopy shows that **3** and **8**—but not control **9**—interact with the BamA β -barrel domain through the external loops L4, L6 and L7 (Fig. 4, Extended Data Figs. 8, 9). The binding interaction also changes the conformational ensemble in the β -barrel lateral gate between open or closed states, and locks BamA in its closed state.

Further work will be necessary to determine how the binding of the chimaeras to BamA causes downstream bactericidal activity. One possibility is that binding inhibits the foldase activity of the BAM complex. The resulting incorrectly folded OMPs, when mislocated to the inner membrane, may lead to cell permeabilization and death²³. Precedence for a link between BamA binding and bactericidal activity comes from a recently described monoclonal antibody (known

as MAB1), which binds an epitope in the external loop L4 on BamA from *E. coli* and causes downstream bactericidal effects²⁴. Previous genetic studies have shown that deletions in the L7 and L8 loops of BamA can lead to severe defects in membrane integrity and outer membrane permeability²⁵, and point mutations in external loops can have marked effects on the activity of the BAM complex²⁶, probably by also influencing conformational changes in BamA that facilitate OMP folding²⁷. However, it is so far unclear whether the chimaeras inhibit BamA function or just use BamA as a binding site. Another possibility is that binding of the chimaeras to BamA provides an additional binding site in the outer membrane that enhances a permeabilizing effect mediated by the polymyxin macrocycle, and helps these antibiotics to avoid LPS-modification resistance mechanisms. On the other hand, the polymyxin macrocycle alone (for example, **6**) (Fig. 1) shows no antimicrobial activity and the mechanism(s) of membrane permeabilization caused by the polymyxins and colistins are presently not known in detail²⁸.

A lead candidate based on these derivatives has—pending future clinical studies—the potential to address life-threatening infections caused by Gram-negative pathogens, and thus to resolve a considerable unmet medical need.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1665-6>.

1. WHO. Global Priority List of Antibiotic-resistant Bacteria to Guide Research, Discovery, and Development of New Antibiotics (World Health Organization, Geneva, 2017).
2. Boucher, H. W. et al. Bad bugs, no drugs: no ESKAPE! An update from the Infectious Diseases Society of America. *Clin. Infect. Dis.* **48**, 1–12 (2009).
3. O'Neill, J. Project Syndicate – A Call to Antimicrobial Arms <https://www.project-syndicate.org/commentary/antibiotics-resistance-economic-costs-by-jim-o-neill-2015-02> (2015).
4. Paterson, D. L. & Harris, P. N. A. Colistin resistance: a major breach in our last line of defence. *Lancet Infect. Dis.* **16**, 132–133 (2016).
5. Henderson, J. C. et al. The power of asymmetry: architecture and assembly of the Gram-negative outer membrane bilayer. *Annu. Rev. Microbiol.* **70**, 255–278 (2016).
6. Kononova, A., Kahne, D. E. & Silhavy, T. J. Outer membrane biogenesis. *Annu. Rev. Microbiol.* **71**, 539–556 (2017).
7. Srinivas, N. et al. Peptidomimetic antibiotics target outer-membrane biogenesis in *Pseudomonas aeruginosa*. *Science* **327**, 1010–1013 (2010).
8. Werneburg, M. et al. Inhibition of lipopolysaccharide transport to the outer membrane in *Pseudomonas aeruginosa* by peptidomimetic antibiotics. *ChemBioChem* **13**, 1767–1775 (2012).
9. Noinaj, N., Rollauer, S. E. & Buchanan, S. K. The β -barrel membrane protein insertase machinery from Gram-negative bacteria. *Curr. Opin. Struct. Biol.* **31**, 35–42 (2015).
10. Storm, D. R., Rosenthal, K. S. & Swanson, P. E. Polymyxin and related peptide antibiotics. *Annu. Rev. Biochem.* **46**, 723–763 (1977).
11. Mares, J., Kumaran, S., Gobbo, M. & Zerbo, O. Interactions of lipopolysaccharide and polymyxin studied by NMR spectroscopy. *J. Biol. Chem.* **284**, 11498–11506 (2009).
12. Roberts, K. D. et al. Antimicrobial activity and toxicity of the major lipopeptide components of polymyxin B and colistin: last-line antibiotics against multidrug-resistant Gram-negative bacteria. *ACS Infect. Dis.* **1**, 568–575 (2015).
13. Baron, S., Hadjadj, L., Rolain, J.-M. & Olaitan, A. O. Molecular mechanisms of polymyxin resistance: knowns and unknowns. *Int. J. Antimicrob. Agents* **48**, 583–591 (2016).
14. Olaitan, A. O., Morand, S. & Rolain, J.-M. Mechanisms of polymyxin resistance: acquired and intrinsic resistance in bacteria. *Front. Microbiol.* **5**, 643 (2014).
15. Raetz, C. R., Reynolds, C. M., Trent, M. S. & Bishop, R. E. Lipid A modification systems in Gram-negative bacteria. *Annu. Rev. Biochem.* **76**, 295–329 (2007).
16. Groisman, E. A. The pleiotropic two-component regulatory system PhoP–PhoQ. *J. Bacteriol.* **183**, 1835–1842 (2001).
17. McPhee, J. B., Lewenza, S. & Hancock, R. E. Cationic antimicrobial peptides activate a two-component regulatory system, PmrA–PmrB, that regulates resistance to polymyxin B and cationic antimicrobial peptides in *Pseudomonas aeruginosa*. *Mol. Microbiol.* **50**, 205–217 (2003).
18. Okuda, S., Sherman, D. J., Silhavy, T. J., Ruiz, N. & Kahne, D. Lipopolysaccharide transport and assembly at the outer membrane: the PEZ model. *Nat. Rev. Microbiol.* **14**, 337–345 (2016).
19. Hartmann, J.-B., Zahn, M., Burmann, I. M., Bibow, S. & Hiller, S. Sequence-specific solution NMR assignments of the β -barrel insertase BamA to monitor its conformational ensemble at the atomic level. *J. Am. Chem. Soc.* **140**, 11252–11260 (2018).

20. Mahoney, T. F., Ricci, D. P. & Silhavy, T. J. Classifying β -barrel assembly substrates by manipulating essential Bam complex members. *J. Bacteriol.* **198**, 1984–1992 (2016).
21. Lee, J. et al. Characterization of a stalled complex on the β -barrel assembly machine. *Proc. Natl Acad. Sci. USA* **113**, 8717–8722 (2016).
22. Gunasinghe, S. D. et al. The WD40 protein BamB mediates coupling of BAM complexes into assembly precincts in the bacterial outer membrane. *Cell Rep.* **23**, 2782–2794 (2018).
23. Mitchell, A. M. & Silhavy, T. J. Envelope stress responses: balancing damage repair and toxicity. *Nat. Rev. Microbiol.* **17**, 417–428 (2019).
24. Storek, K. M. et al. Monoclonal antibody targeting the β -barrel assembly machine of *Escherichia coli* is bactericidal. *Proc. Natl Acad. Sci. USA* **115**, 3692–3697 (2018).
25. Browning, D. F. et al. Mutational and topological analysis of the *Escherichia coli* BamA protein. *PLoS ONE* **8**, e84512 (2013).
26. Lee, J. et al. Substrate binding to BamD triggers a conformational change in BamA to control membrane insertion. *Proc. Natl Acad. Sci. USA* **115**, 2359–2364 (2018).
27. Rigel, N. W., Ricci, D. P. & Silhavy, T. J. Conformation-specific labeling of BamA and suppressor analysis suggest a cyclic mechanism for β -barrel assembly in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **110**, 5151–5156 (2013).
28. Dixon, R. A. & Chopra, I. Polymyxin B and polymyxin B nonapeptide alter cytoplasmic membrane permeability in *Escherichia coli*. *J. Antimicrob. Chemother.* **18**, 557–563 (1986).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

MICs

Clinical isolates (collected between 2012 and 2017) were obtained from the University Hospital Basel and from the IHMA collection (IHMA Europe Sarl). *E. coli* strains containing the *mcr-1* gene were obtained from P. Nordmann. LPS-deficient strains of *A. baumannii* were obtained from J. Moffatt. Reference strains were obtained from the American Type Culture Collection (ATCC), the Deutsche Sammlung für Mikroorganismen und Zellkulturen (DSMZ) and the National Collection of Type Cultures (NCTC). The strains were stored at -80°C as 30% (v/v) glycerol cultures. Large numbers of colistin-resistant isolates, as well as isolates resistant to carbapenems and third-generation cephalosporins, were included in the test panels. All isolates were tested by the CLSI broth microdilution method (M07-A10) in cation-adjusted MH broth (CAMHB), and the EUCAST interpretive criteria were used to determine susceptibility for comparators. The compounds were tested in sterile 96-well microtitre plates in the range from 8 to 0.0078 mg l^{-1} in the presence of polysorbate-80 (P-80 or Tween-80, sterile-filtered) at 0.002% v/v final concentration. For comparison, the commercial antibiotics were tested in the range from 64 to 0.06 mg l^{-1} without P-80, according to the CLSI guidelines. In some experiments, MICs were determined in MH broth, non-cation-adjusted. MICs in serum were measured by addition of pooled human serum (Blutspendezentrum Basel) to a final concentration of 50% (v/v) in CAMHB. Supplementary Table 2 provides: (a) MIC at which at least 50% of the isolates were inhibited (MIC_{50}), MIC at which at least 90% of the isolates were inhibited (MIC_{90}) and ranges (mg l^{-1}) against *A. baumannii* (30 isolates, 57% multidrug resistant), *Enterobacter* spp. (28 isolates), *E. coli* (28 isolates, 30% multidrug resistant), *K. pneumoniae* (31 isolates, 50% multidrug resistant) and *P. aeruginosa* (29 isolates, 31% multidrug resistant); (b) MICs (mg l^{-1}) with and without 50% human serum against representative Gram-negative strains; (c) MICs (mg l^{-1}) against selected LPS-less *A. baumannii* mutant (*lpxA* deletion) compared to wild-type parent strain²⁹; and (d) MICs (mg l^{-1}) of **4** and **8**, and their enantiomers (**e-4** and **e-8**, respectively), against representative Gram-negative strains and *S. aureus*.

Kill-curve kinetics

MIC values were determined by a macrodilution method in 6-well plates (3 ml) in CAMHB. After 18–22 h at 35°C , the first well without growth was taken as the MIC value used for kill-curve experiments. Time-kill assays were done in tubes (in 1.5 ml of CAMHB) and antibiotics at 1 \times , 2 \times and 4 \times the MIC. Tubes were inoculated with bacteria from exponentially growing cultures and incubated at 35°C under shaking. At 0, 2, 4, 8 and 24 h, CFUs were enumerated on MH agar plates. The concentration and time to reach a $\geq 3\log_{10}$ reduction with no re-growth at 24 h was recorded. Results are shown in Extended Data Fig. 1.

Serial passage

Resistance development during serial passage exposure was assessed in tubes containing 1 ml of twofold dilutions of antibiotics in CAMHB. Tubes were inoculated 10% v/v with a bacterial suspension adjusted to 0.5 McFarland standard. Tubes were incubated for 24 h at 35°C under shaking. After each passage, the tube with the lowest drug concentration showing no growth was recorded as the passage (tube) MIC, and the tube with the highest drug concentration showing substantial growth was used to inoculate the tubes of the next passage (1% of a suspension adjusted to 0.5 McFarland standard). Experiments were stopped after 21 serial passages or when the passage MIC exceeded the resistance breakpoint for 3 consecutive passages. At selected passages and at the end of the experiments, samples from tubes showing bacterial growth at the highest antibiotic concentration were plated on MH agar plates. Single colonies were isolated, passaged once on drug-free agar and subjected to standard MIC testing against chimeric and reference antibiotics (see 'MICs').

Haemolysis

To test their haemolytic potential, compounds were incubated in the presence of phosphate-buffered-saline (PBS)-washed human red blood cells (Blutspendezentrum Basel). After 1 h incubation at 200 mg l^{-1} and 37°C , the samples were centrifuged at 3,220g and the supernatants were diluted in Dulbecco's PBS (DPBS) followed by spectrophotometric measurement (optical density at 540 nm (OD_{540})). The haemolysis induced by the compound was calculated versus a 100% lysis control prepared with 2.5% Triton X-100. Biological replicates were used for haemolysis determination (Supplementary Table 3).

Cytotoxicity

Cytotoxicity of compounds was determined using WST-8 (Sigma Cell Counting Kit-8). Exponentially growing HeLa and HEP G2 cells were seeded in 96-well microtitre plates in appropriate cell culture medium. After 24 h incubation at 37°C and 5% CO_2 , medium was replaced with fresh, phenol-red-free medium containing dilutions of compounds. Maximum assay concentrations were 100 mg l^{-1} for HeLa cells and 200 mg l^{-1} for HEP G2-cells. Following 48 h incubation, cell viability was monitored by addition of WST-8 solution and measurement of optical density at 450 nm (OD_{450}) after 1 h. The experiments were performed using biological replicates (Supplementary Table 3).

Plasma stability

Compounds were incubated in K_3EDTA -stabilized plasma (mixed gender) of human and CD-1 mouse (Seralabs). Their stability was assessed in triplicate at 10 mg l^{-1} and after an incubation at 37°C . Samples were taken at 0, 15, 30, 60, 120 and 240 min and extracted by precipitation with 3 volumes of acetonitrile + 0.5% TFA. Sample quantitative analysis was performed by high-performance liquid chromatography (HPLC) with tandem mass spectrometry (MS/MS) (Supplementary Table 3).

Protein binding

The binding of compounds to proteins in pooled human and CD-1 mouse K_3EDTA -stabilized plasma (mixed gender) was determined by ultrafiltration method using a 30-kDa cut-off filter in filter tubes (Millipore Centrifree). Compounds were diluted in pH 7.5-adjusted plasma to a final concentration of 10 mg l^{-1} and incubated for 30 min at 37°C . After incubation, the unbound fraction (f_u) was separated by ultrafiltration. Protein binding was determined by subtracting the percentage of compound in ultrafiltrate (that is, the f_u) from the total amount of compound in spiked plasma (Supplementary Table 3).

In vivo tolerability studies

The 'Autonomic Signs' study design was used to determine the tolerability of compounds after single or multiple dosing. The goal was to assess a dose that does not produce mortality or overt clinical signs of toxicity to be used for pharmacokinetics studies and in vivo efficacy. Test substances were administered subcutaneously to a group of three male ICR mice. The mice were observed for the presence of acute toxic symptoms and autonomic effects during the first 30 min. The mice were then observed again for mortality at 3, 24, 48 and 72 h after compound administration (Supplementary Table 3).

Pharmacokinetic analysis

Adult CD-1 male mice were injected subcutaneously with a single dose of 10 mg/kg of a 0.9% saline test item formulation adjusted to pH 6.5–7.6. Plasma samples were taken from 9 mice for single administration in each treatment group at 0.25, 0.5, 1, 2, 3, 4, 8 h post-dose. Blood samples (in Li-heparin as anticoagulant) were collected from the retrobulbar venous plexus under short isoflurane anaesthesia. Plasma samples were obtained by centrifugation for 10 min at 3,000g and 4°C . The protein-free supernatant was analysed by LC–MS/MS using a Q Exactive hybrid quadrupole Orbitrap mass spectrometer coupled with an Accela UHPLC

Article

system and an AS Open autosampler (Thermo Fisher Scientific). After separation on an Accucore phenyl–hexyl reverse phase column using an acetonitrile–water gradient, peaks were analysed by mass spectroscopy using electrospray ionization. The mean plasma concentration and the standard deviation from all three mice within each time point were calculated, and pharmacokinetics parameters of test agent were calculated with a non-compartmental analysis model based on WinNonlin, using a trapezoid area calculation (Supplementary Table 3).

Mouse model of systemic infection

To evaluate the in vivo efficacy of peptide **3**, adult male CD-1 mice (6 per group) were infected by intraperitoneal administration of 9.2×10^5 CFU/ml *E. coli* 5799. The peptide **3** was dosed at 6.25, 3.13 and 1.56 mg/kg, at 1 h post-infection. The survival at day 7 (%) was recorded (Extended Data Fig. 1d).

Mouse models of peritonitis

To evaluate the in vivo efficacy of peptides **3** and **8**, adult male CD-1 mice (6 per group) were rendered neutropenic with injections with cyclophosphamide on day –4 and day –1. On day 0, mice were infected by intraperitoneal administration of either 1.4×10^7 CFU/ml *E. coli* AF45 (*mcr-1*), 1.4×10^7 CFU/ml of *K. pneumoniae* SSI3010 or 3.3×10^7 CFU/ml *E. coli* SNTR36B6 (*mcr-3*). The CFUs in the blood and/or in the peritoneal wash were counted at the start of treatment and at the end of treatment. The compounds were dosed at 30 mg/kg (*E. coli* AF45 (*mcr-1*)) and 10 mg/kg (*K. pneumoniae* SSI3010 and *E. coli* SNTR36B6 (*mcr-3*)) and the control antibiotics for the studies were for these three strains tigecycline (40 mg/kg), ciprofloxacin (13 mg/kg) or meropenem (40 mg/kg), respectively (Extended Data Fig. 1e–h).

Mouse models of thigh infection

To evaluate the in vivo efficacy of **8**, adult male CD-1 mice (6 per group) were rendered neutropenic with injections with cyclophosphamide on day –4 and day –1. Mice were infected 24 h after the second dose of immunosuppressive agent by intramuscular instillation of bacterial inocula (about 2×10^7 CFU/ml, corresponding to about 1×10^6 CFU per thigh). Treatments were administered twice in total, at 2 and 14 h after infection. Additional groups were included that were euthanized pre-treatment (2 h after infection) or treated with vehicle only. The vehicle control group was treated with 0.9% saline, also at 2 h and 14 h after infection. At 26 h post-infection, the clinical condition of all mice was assessed and the mice were humanely euthanized by pentobarbitone overdose (Extended Data Fig. 1i–k). Mouse weight was determined before the thighs were removed and weighed. Thigh-sample homogenates were quantitatively cultured onto agar for determination of the counts of CFU per thigh.

Mouse renal toxicity

Adult male CD-1 mice (5 per group) were dosed subcutaneously at 12 mg/kg, 6 times a day (every 2 h) and clinical signs were monitored (Supplementary Table 3). At termination, all mice were subjected to gross necropsy and tissues were examined macroscopically. The kidneys were histopathologically examined, semiquantitative scoring of the kidneys was performed and lesions were rated as follows: mild acute tubular damage with tubular dilation, prominent nuclei and a few pale tubular casts (grade 1); severe acute tubular damage with necrosis of tubular epithelial cells and numerous tubular casts (grade 2); and necrosis and/or infarction of tubules and glomeruli, with or without papillary necrosis (grade 3). Subsequently, the overall kidney histology score was calculated as the product of percentage score and grade score. These scores were then expressed as a semiquantitative score (SQS) on a scale of 0 to 5 for renal histological changes. These scores were assigned as follows: SQS 0, no substantial change (overall score, <1); SQS 1, mild damage (overall score, 1 to <15); SQS 2, mild-to-moderate damage (overall score, 15 to <30); SQS 3, moderate damage (overall score, 30 to <45); SQS 4,

moderate-to-severe damage (overall score, 45 to <60); and SQS 5, severe damage (overall score, >60).

Photo-affinity interaction mapping

E. coli ATCC 25922 cells grown in MH-II broth (50 ml) to an optical density at 600 nm (OD_{600}) of 1.0 were collected, washed once and taken up in PBS (50 ml) and incubated for 30 min at 37 °C with shaking at 200 r.p.m. in the dark with 4–10 µg/ml photoprobe. Photo-activation was achieved by UV irradiation at 350 nm in a Rayonet Reactor (16 × 8 W Sylvania blacklight lamps) for 30 min at 30 °C. Cells were then collected and washed twice with PBS. Cell pellets were stored at –20 °C. The cell pellet was resuspended in PBS, with protease inhibitor cocktail (cOmplete, Roche) and lysed by 3 cycles of sonication using a Branson digital sonifier equipped with a microtip (80 W, 30% intensity, 20 s on with 20 s off for 2 min) under cooling on ice. To remove unbroken cells and cell debris, the lysate was centrifuged (20 min at 4,000 r.p.m., 4 °C). The supernatant was subjected to ultracentrifugation (45,000g, in a Sorvall T-875 rotor, 1 h, 4 °C). The pellet was washed with PBS and collected again by ultracentrifugation (1 h, 4 °C).

Membrane protein fractions in SDS loading buffer containing DTT (100 mM) were boiled for 5 min at 100 °C, before SDS–PAGE under standard Laemmli conditions. Proteins were blotted to PVDF membrane (0.45-µm pore size, Immobilon-P, Merck) using a 1:1 mixture of Tris–glycine–SDS (12 mM Tris, 96 mM glycine, 0.1% SDS) and phosphate–SDS–urea buffer (10 mM Na₂HPO₄, 1% SDS, 6 M urea). Blotting of proteins from gel onto the membrane was achieved using a Pierce G2 Fast Blotter (Thermo Fisher) for 2 h at 0.5 A and 10 V. For chemiluminescence detection, the blocked membrane was incubated with neutravidin–HRP conjugate (Pierce, diluted 1:30,000 in PBS, 1% BSA, 0.2% Tween-20) for 1 h. The membrane was washed 4 × 5 min with PBS and developed with WesternBright Sirius (Advanta) HRP substrate. Chemiluminescence was detected on a ChemoDoc MP Imaging System (Bio-Rad) over the course of 1–10 min. The results are shown in Extended Data Fig. 6.

E. coli cells were photolabelled with **PAL-3** ($n = 3$), **PAL-4** ($n = 3$) or **PAL-7** ($n = 3$), or treated the same way with **3** ($n = 4$), **4** ($n = 4$) or **7** ($n = 3$) with n being biologically independent samples. Cells were washed with PBS and lysed in ice-cold 50 mM (NH₄)HCO₃ (AmBic) containing protease inhibitor cocktail (Roche, 11704900) and 0.1% RapiGest (Waters, 186002122) by 4 intervals of 15-s ultrasound sonication in a vial tweeter (Hielscher Ultrasonics) at a power of 170 W and 80% cycle time. Protein concentration was determined using a Nanodrop 2000 Spectrophotometer (Thermo Fisher Scientific) and 15 mg protein of each sample was subjected to automated purification and processing of biotinylated proteins. For this, in-house-packed tips containing 80 µl Streptavidin Plus UltraLink resin (Thermo Fisher Scientific) were linked to a Versette liquid handling robotic system (Thermo Fisher Scientific) and incubated with the cell lysate for 2.5 h at room temperature by pipetting up and down. In an automated fashion, bead-bound proteins were subsequently washed with 5 M NaCl, StimLys buffer (50 mM Tris pH 7.8, 137 mM NaCl, 150 mM glycerol, 0.5 mM EDTA, 0.1% Triton X-100), 100 mM NaHCO₃ and AmBic, reduced with 5 mM Tris(2-carboxyethyl)phosphine (TCEP) in 3 M urea and AmBic for 30 min at 37 °C and alkylated with 10 mM iodoacetamide in 3 M urea and AmBic for 30 min at 37 °C. Bead-bound proteins were proteolytically digested with 0.5 µg lysyl endopeptidase Lys-C (Wako, 125-05061) in 3 M urea and AmBic for 2 h at 37 °C before diluting to 1.5 M urea and AmBic, and adding 0.8 µg sequencing grade modified trypsin for 14 h at 37 °C. Eluted peptides were acidified to pH <3 by the addition of formic acid and subjected to C18 purification using 5–60 µg UltraMicroSpin Columns (The Nest Group, SEM SS18V) according to the manufacturer's instructions.

Peptide samples were separated by reversed-phase chromatography on a HPLC column (75-µm inner diameter, New Objective) that was packed in-house with a 15-cm stationary phase (ReproSil–Pur C18–AQ, 1.9 µm) and connected to a nano-flow HPLC with an autosampler (EASY-nLC

1000, Thermo Scientific). The HPLC was coupled to a Q-Exactive plus mass spectrometer (Thermo Scientific) equipped with a nano electrospray ion source (Thermo Scientific). Peptides were loaded onto the column with 100% buffer A (99.9% H₂O, 0.1% FA) and eluted at a constant flow rate of 300 nL/min with a 70-min linear gradient from 6–28% buffer B (99.9% MeCN, 0.1% FA) followed by a 4-min transition from 28 to 50% buffer B. After the gradient, the column was washed for 10 min with 98%, 4 min with 10% and again 8 min with 98% buffer B. Electrospray voltage was set to 2.2 kV and capillary temperature to 250 °C. In data-dependent acquisition mode, the mass spectrometer automatically switched between precursor- and fragment-ion detection. Following a high-resolution survey mass spectrum (from 300 to 1,700 *m/z*) acquired in the Orbitrap with resolution *R* = 70,000 at *m/z* 200 (automatic gain control target value 3×10^6), the 15 most-abundant peptide ions with a minimum intensity of 2.5×10^4 were selected for subsequent higher-energy collision-induced dissociation fragmentation, with an isolation window of 1.4 Da, and fragments were detected by MS/MS acquisition in the Orbitrap at resolution *R* = 35,000 (automatic gain control target value 1×10^6). Target ions already selected for fragmentation were dynamically excluded for 30 s. Acquired raw files were subjected to protein identification using Comet (v.2015.01) and TransProteomic Pipeline v.4.7 (SPC/ISB Seattle) by matching ion spectra acquired in data-dependent acquisition mode against a SwissProt (UniProt consortium) reviewed *E. coli* protein database (downloaded November 2016) containing common contaminants. Peptides were required to be fully tryptic with a maximum of two missed cleavage sites, carbamidomethylation as fixed modification and methionine oxidation as a dynamic modification. The precursor and fragment mass tolerance were set to 20 ppm and 0.02 Da, respectively. Proteins identified by at least two proteotypic peptides were quantified by integration of chromatographic traces of peptides using Progenesis Q1 v.4.0 (Nonlinear Dynamics). Contaminant hits were removed and proteins filtered to obtain a false discovery rate of < 1%. Raw protein abundances were exported based on non-conflicting peptides. Within the R computing environment, protein abundance changes (expressed in log₂) were calculated by linear mixed-effect model and tested for statistical significance using a two-sided *t*-test by using the R package MSstats v.3.5.3³⁰. The *P* values obtained were further corrected for multiple comparisons using Benjamini–Hochberg method. Subcellular-localization annotation was retrieved from UniProt database using R package UniProt.ws v.2.14.0³¹. Significantly enriched, outer-membrane-annotated *E. coli* proteins (abundance fold change ≥ 1.5 and adjusted *P* ≤ 0.05) were considered to be bona fide **PAL-3**-, **PAL-4**- or **PAL-7**-interacting candidates. Mass spectrometry data are available at the ProteomeXchange Consortium (<http://www.proteomexchange.org/>) with the dataset identifier PXD010174.

Interaction-site mapping with BamA^{ext}

The available sequence-specific resonance assignments of BamA^{ext} have recently been reported and deposited in the BMRB database, with accession code 27431¹⁹. For NMR binding studies, [*U*-¹⁵N, ²H]-labelled BamA^{ext} in LDAO micelles was prepared in 20 mM HEPES pH 7.5, 150 mM NaCl and 0.1% LDAO. Spectra were recorded on a Bruker Avance-700

spectrometer equipped with a cryogenic probe. Two-dimensional [¹⁵N, ¹H]TROSY spectra were acquired at 37 °C with 24 or 80 scans and 128 and 1024 complex points in the δ₁(¹⁵N) and δ₂(¹H) dimension, respectively. Peptides were titrated from a stock solution to the protein. The chemical shifts of the titrated form were tracked by a stepwise titration, which was—in most cases—possible without ambiguity. Combined chemical shift perturbations upon peptide addition for amide moieties were calculated as $\Delta\delta(\text{HN}) = \sqrt{(\Delta\delta(^1\text{H}))^2 + (0.2 \times \Delta\delta(^{15}\text{N}))^2}$. Residues were considered to be significant interactors when they had a chemical shift perturbation of $\Delta\delta(\text{HN}) > 0.07$ ppm or if their signal intensity vanished owing to chemical exchange upon peptide addition (Extended Data Fig. 8).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Mass spectrometric data are available at the ProteomeXchange Consortium (<http://www.proteomexchange.org/>) with dataset identifier PXD010174.

29. Moffatt, J. H. et al. Colistin resistance in *Acinetobacter baumannii* is mediated by complete loss of lipopolysaccharide production. *Antimicrob. Agents Chemother.* **54**, 4971–4977 (2010).
30. Choi, M. et al. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **30**, 2524–2526 (2014).
31. Carlson, M. UniProt.ws: R interface to UniProt web services. R package version 2.14.0. <https://bioconductor.org/packages/release/bioc/html/UniProt.ws.html> (2018)

Acknowledgements We thank M. Gwerder, the Center for Microscopy and Image Analysis at UZH, and the Functional Genomics Center Zurich for technical support. We thank the following agencies for funding: M. Müller and B.W. were supported by ETH grant ETH-30 17-1 and a grant from the Swiss National Science Foundation (grant number 31003A_160259); J.A.R. was supported by a grant from the Commission for Technology and Innovation (CTI/KTI) (grant number 18146.1 PFLS-LS); S.H. was supported by a grant from the Swiss National Science Foundation via the NRP 72 (grant 407240_167125); A.V. was supported by the Swiss National Science Foundation (SystemsX.ch - IPhD project 51PHPO_163556). Polyphor acknowledges funding from CARB-X and the Wellcome Trust (grant number 202728/Z/16/Z) and financial support from the REPAIR Impact Fund (Novo Holdings).

Author contributions F.B., A. Luther, A. Lederer, G.E.D., P.C., S.S., C.V., T.R., A.W., P.R., S.M.M., M.S., C.K., M.-A.W., N.D., E.B., S.H., K.L., A.V., R.J., V.R., G.U., P.Z., H.H.L. and D.O. performed discovery chemistry and biological evaluations; M.Z., T.S., J.-B.H. and S.H. conceived and performed NMR and binding studies; K.Z., M.U., M. Mondal, S.-Y.W., F.L.M., E.C., H.K., K.M. and J.A.R. conceived and performed mechanism-of-action analyses; M. Müller and B.W. conceived and performed mass-spectrometry-based proteomic studies; A.V., G.P. and L.E. performed molecular genetic and microbiological studies. All authors contributed to the analysis and interpretation of results, and J.A.R. and D.O. wrote the paper, which was seen and agreed by all authors.

Competing interests A. Luther, P.C., S.S., C.V., T.R., M.S., C.K., M.-A.W., N.D., E.B., S.H., K.L., A.V., R.J., V.R., G.U., A. Lederer, P.Z., A.W., H.H.L., F.B., G.E.D. and D.O. declare competing interests as employees of Polyphor AG who pursue clinical studies.

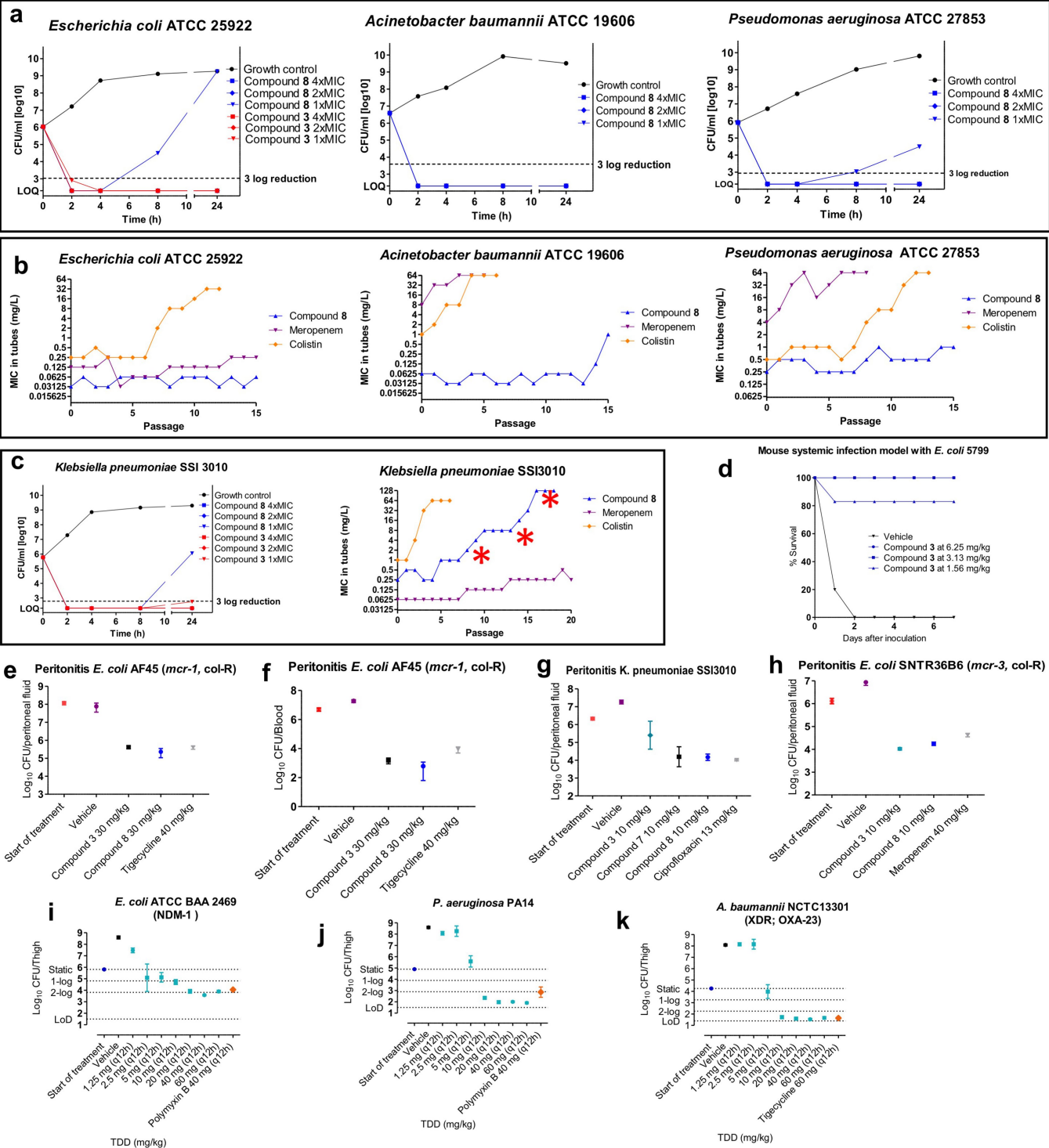
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1665-6>.

Correspondence and requests for materials should be addressed to J.A.R. or D.O.

Peer review information Nature thanks Paul Hergenrother, Lynn Silver and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

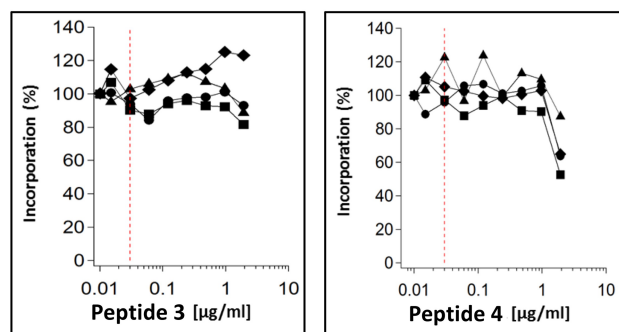


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Biological properties of the chimeric antibiotics.

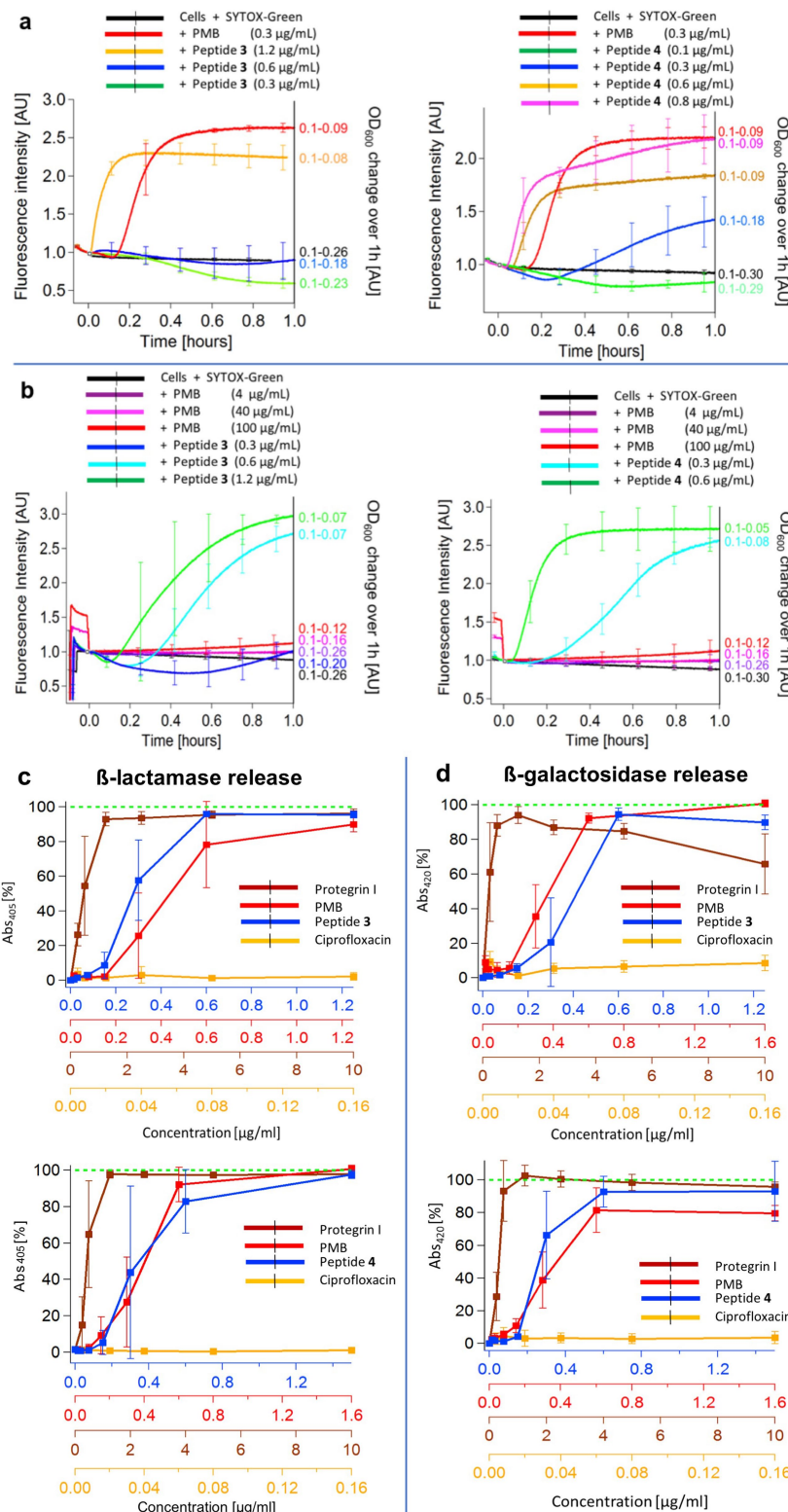
a, In vitro killing kinetics of **3** and **8** against representative Gram-negative species, *E. coli* ATCC 25922, *P. aeruginosa* ATCC 27853 and *A. baumannii* ATCC 19606. **b**, Resistance development of **8** by serial passage against *E. coli* ATCC 25922, *P. aeruginosa* ATCC 27853 and *A. baumannii* ATCC 19606. The y axis indicates the MIC measured directly from the tubes during the serial passages (mg l^{-1}) and the x axis is the number of passages. Antibiotics used as comparisons are indicated (colistin and meropenem). In **a**, **b**, the curves show typical examples of $n = 2$ biologically independent experiments. **c**, In vitro killing kinetics of **3** and **8** against *K. pneumoniae* SSI3010 and resistance development of **8** by serial passage against *K. pneumoniae* SSI3010. The y axis indicates the MIC measured directly from the tubes during the serial passages (mg l^{-1}) and the x axis is the number of passages. The asterisks represent the clones taken for whole-genome sequencing. Antibiotics used as comparisons are indicated (colistin and meropenem). The curves show representative examples of $n = 2$ biologically independent experiments. **d**, In vivo efficacy of peptide **3** in a mouse model of septicemia against *E. coli* 5799. Each point represents the

percentage of survival of $n = 6$ mice. **e**, **f**, In vivo efficacy of **3** and **8**, in mouse models of peritonitis, against *E. coli* AF45 (*mcr-1*, colistin resistant) after a single subcutaneous administration (reduction in CFU counts in peritoneal wash fluid and blood). The mean and s.e.m. of $n = 6$ mice are shown. **g**, In vivo efficacy of **3**, **7** and **8**, in mouse models of peritonitis, against *K. pneumoniae* SSI3010 after a single subcutaneous administration (reduction in CFU counts in peritoneal wash fluid). The mean and s.e.m. of $n = 4$ ($n = 6$ for vehicle) mice are shown. Start of treatment and vehicle were repeated in three experiments. **h**, In vivo efficacy of **3** and **8**, in mouse models of peritonitis, against *E. coli* mcr-3 SNT R36B6 (colistin resistant) after a single subcutaneous administration (reduction in CFU counts in peritoneal wash fluid). The mean and s.e.m. of $n = 4$ mice ($n = 6$ for vehicle) are shown. **i–k**, In vivo efficacy of **8**, in mouse models of thigh infection, against *E. coli* ATCC BAA2469 (NDM-1 strain), *P. aeruginosa* PA14 and *A. baumannii* NCTC 13301 (extensively drug resistant, OXA-23 strain). The total daily dose (TDD) indicated was administered in 2 doses over the course of 24 h (q12h). The mean and s.e.m. of $n = 6$ mice are shown ($n = 4$ mice for start of treatment). On each mouse, two technical replicates were done.



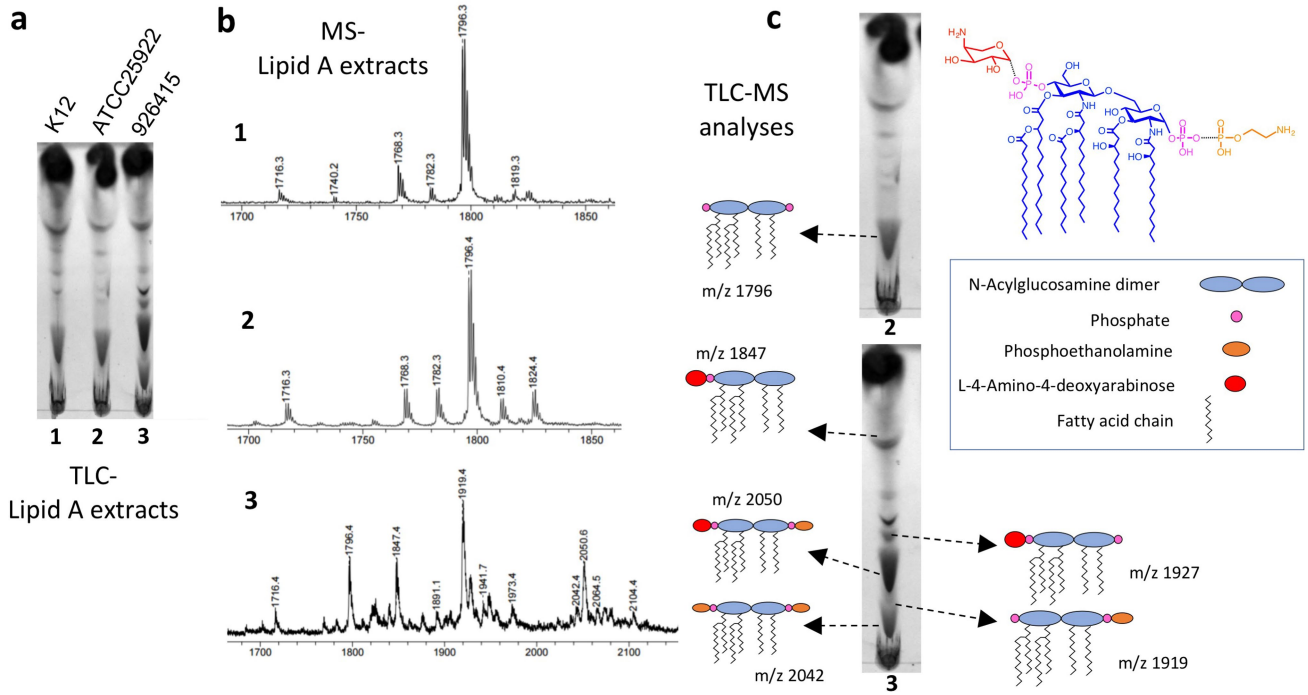
- Thymidine[Me-³H]
- ▲ Uridine[5,6-³H]
- L-Leucine[4,5-³H]
- ◆ N-Acetyl-D-glucosamine [1,6-³H]

Extended Data Fig. 2 | Macromolecular synthesis assays. Relative incorporations of ³H label into macromolecules in *E. coli* from the labelled precursors shown over 20 min, at 37 °C, with increasing concentrations of peptide 3 and peptide 4. The incorporation is relative to a control with no addition of antibiotic (100%). The results show no inhibition of protein, RNA, DNA biosynthesis or cell-wall polysaccharide. In control experiments, the expected effects of known antibiotics (tobramycin, rifampicin, ciprofloxacin or ceftriaxone, on protein, DNA, RNA or cell-wall biosynthesis, respectively, were observed (data not shown)). The red dotted line indicates the MIC of each antibiotic. The results shown are representative of *n* = 3 biologically independent experiments.



Extended Data Fig. 3 | Permeabilization assays. a, b. Membrane permeabilization elicited by **3**, **4** or PMB monitored by uptake of SYTOX-Green, and increase in fluorescence intensity. **a**, Permeabilization with *E. coli* ATCC 25922. **b**, Permeabilization with PMB-resistant *E. coli* 926415 strain. No fluorescence increase is seen from cells in the presence of SYTOX-Green without antibiotic. The change in cell density (OD₆₀₀) in the cuvette over 1 h is shown on the right. For experimental methods, see Supplementary Information. In **a**, the curves represent the mean of $n = 3$ biologically independent experiments (except for PMB on the left, and peptide **4** (0.8 $\mu\text{g mL}^{-1}$), for which $n = 2$ biologically independent experiments). The error bars are s.d.

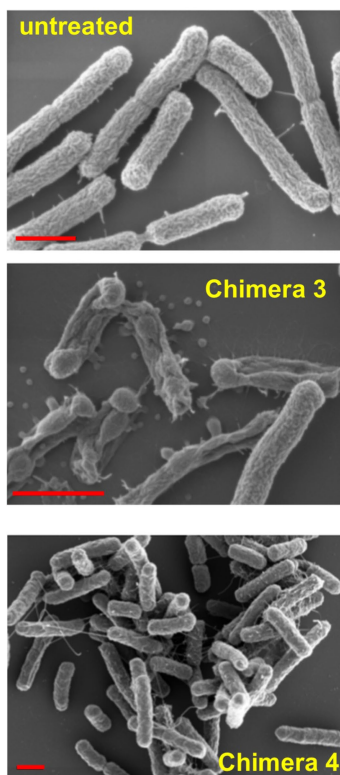
In **b**, the curves represent the mean of $n = 3$ biologically independent experiment (except for PMB (40 $\mu\text{g mL}^{-1}$), peptide **3** (0.3 $\mu\text{g mL}^{-1}$) and peptide **4** (0.3 $\mu\text{g mL}^{-1}$), for which $n = 2$ biologically independent experiments). The error bars are s.d. **c, d**, Release from *E. coli* of β -lactamase (**c**) and of β -galactosidase (**d**) in the presence of **3**, **4** or PMB, monitored by enzymatic assays. Ciprofloxacin does not cause detectable release of either enzyme from cells, whereas protegrin I caused rapid and complete release of both enzymes (100% value corresponds to enzyme released by sonication of cells). For experimental methods, see Supplementary Information. The curves represent the mean of $n = 3$ or 4 biologically independent experiment. The error bars show the s.d.



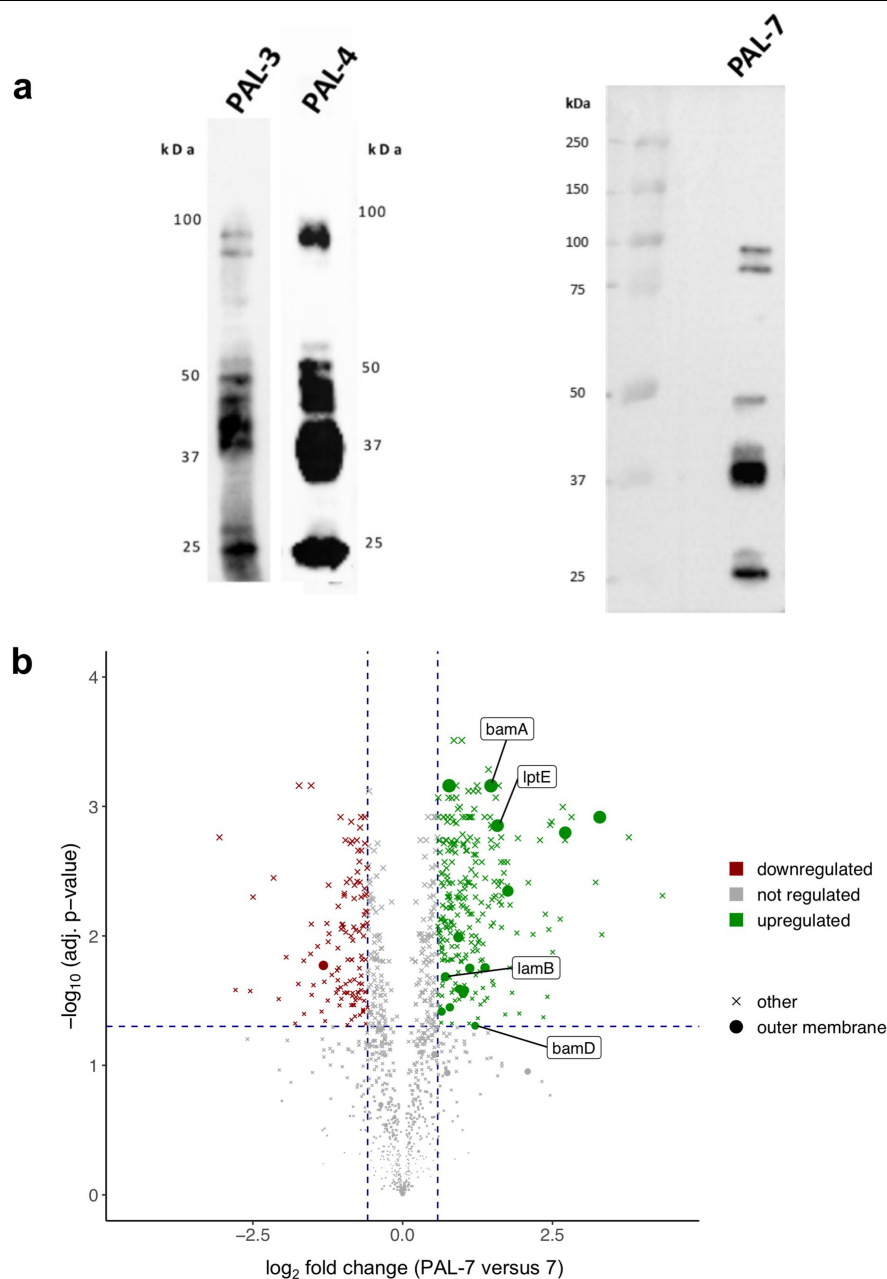
Extended Data Fig. 4 | Analysis of lipid A. Thin layer chromatography (TLC) and mass spectrometry analyses of lipid A. **a**, TLC analysis of lipid A extracted from *E. coli* K12 (**1**), ATCC 25922 (**2**) and PMB-resistant strain 926415 (**3**).

b, Matrix-assisted laser desorption/ionization–time of flight (MALDI-TOF) mass spectrometry of the lipid A mixture isolated from each strain. **c**,

Individual lipid A species identified in the lipid A extracted from strains ATCC 25922 and 926415 by TLC-MALDI-TOF mass spectrometry. All lipid A in sample 3 contains phosphoethanolamine and/or L-4-amino-4-deoxyarabinose units. For experimental methods, see Supplementary Information. The results shown are representative of $n = 2$ biologically independent experiments.

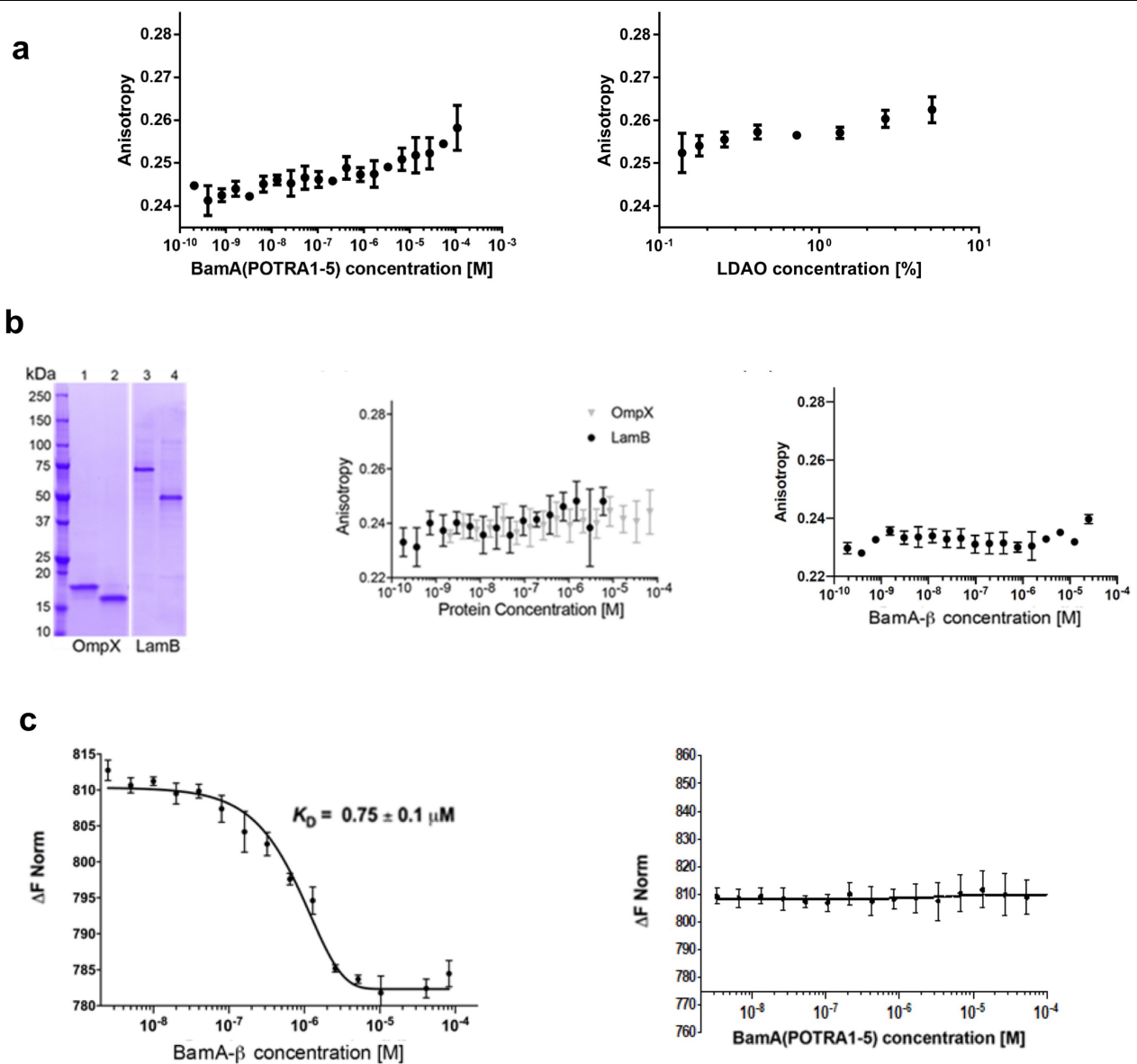


Extended Data Fig. 5 | Scanning electron microscopy. Scanning electron microscopy of *E. coli* ATCC 25922 cells untreated or grown with **3** or **4**, at concentrations that cause a growth inhibition of about 50% (about 0.1 mg l^{-1}). Scale bars, $1 \mu\text{m}$. The scanning electron microscopy scans were in completed in duplicate, and one typical result is shown. For experimental methods, see Supplementary Information.



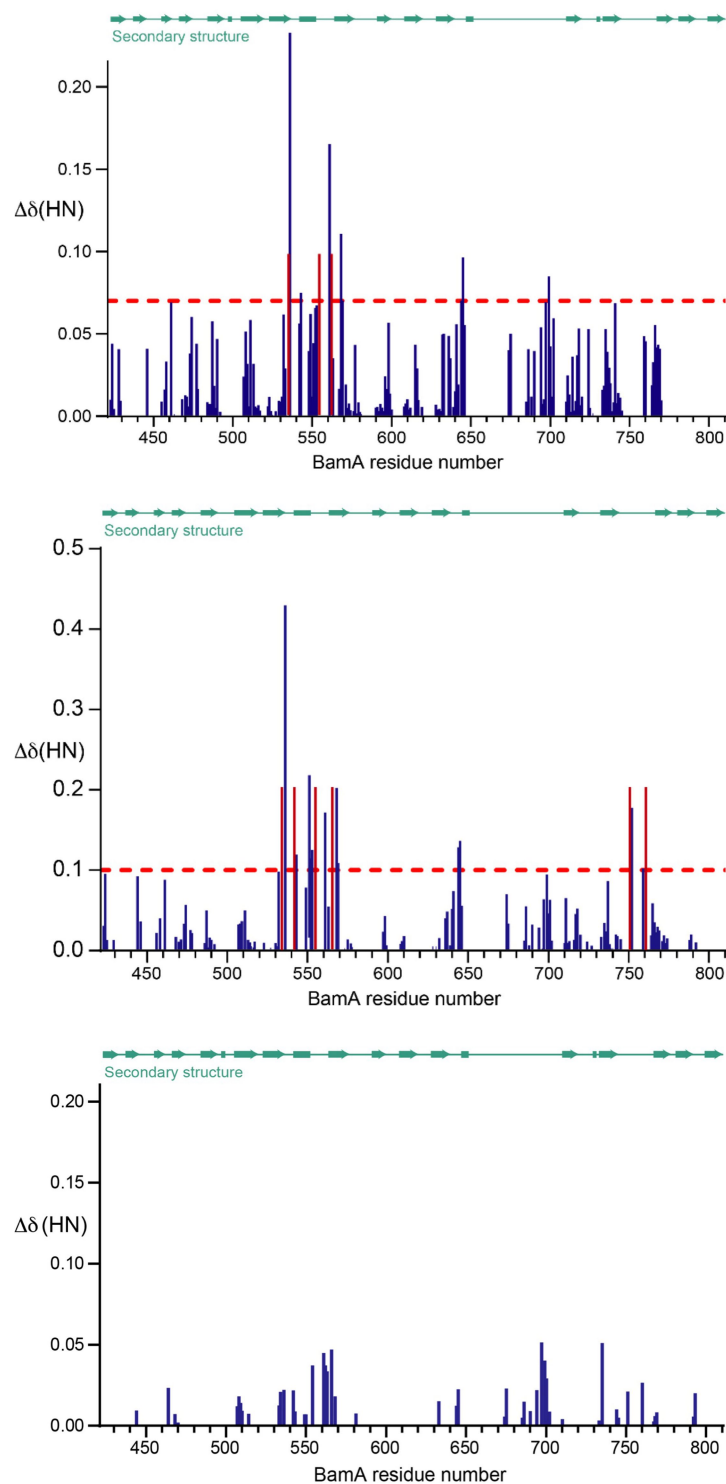
Extended Data Fig. 6 | Photo-affinity interaction mapping. **a**, Western blots (10% SDS-PAGE gel, blotted to a PVDF membrane) of membrane protein extract from **PAL-3** (4 mg l⁻¹), **PAL-4** (10 mg l⁻¹) and **PAL-7** (4 mg l⁻¹) labelled *E. coli* ATCC 25922 with chemiluminescence detection of biotinylated macromolecules. For gel source data, see Supplementary Fig. 1. **b**, Volcano plot showing relative abundance of proteins captured by streptavidin, and quantified by mass spectrometry, from *E. coli* cells photolabelled with **PAL-7** versus control cells treated with 7 ($n=3$ biologically independent samples each). Protein abundance changes (expressed in \log_2) were calculated by linear mixed-effect model and tested for statistical significance using a two-sided *t*-test. *P* values obtained were further corrected for multiple comparisons using Benjamini–

Hochberg method. Proteins are represented on the basis of the UniProt annotated subcellular location as dots (outer membrane) or crosses (no, or other, location); symbol size is scaled according to statistical significance. Significantly enriched proteins (abundance ratio ≥ 1.5 and adjusted $P \leq 0.05$, shown as blue lines) are coloured in green. Outer membrane proteins that were also enriched in **PAL-3** and **PAL-4** photo-affinity interaction mapping experiments are highlighted. BamA is among the most significantly upregulated proteins, and is the only common outer membrane interaction candidate that was identified by all three photoprobes. A full list of proteins quantified by mass spectrometry in these experiments is supplied as Source Data.



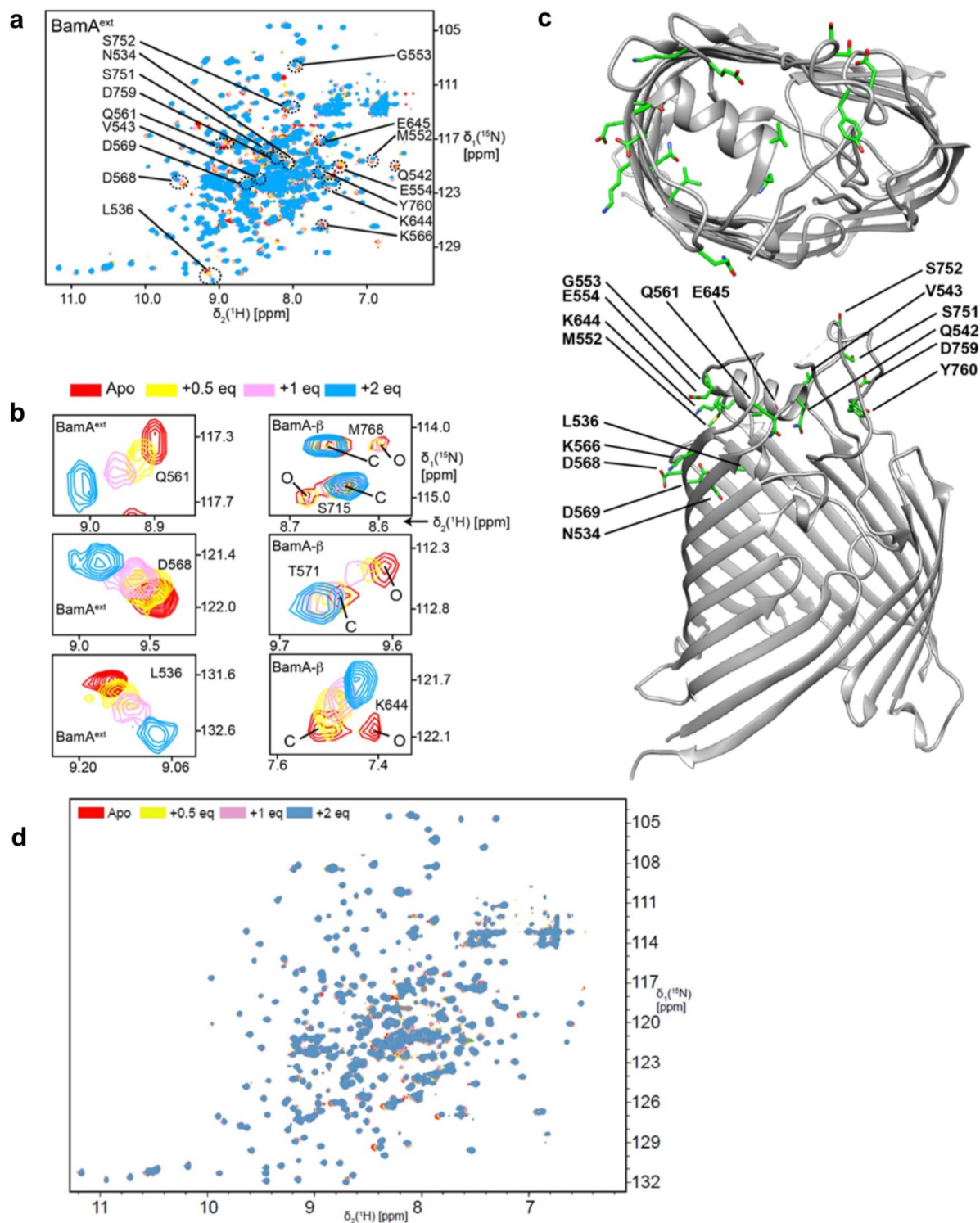
Extended Data Fig. 7 | Binding assays. **a**, Left, titration of BamA(POTRA1-5) to peptide **Cy3-3**, monitored by fluorescence anisotropy. Measurements were performed in triplicates; s.d. around the mean is shown. No interaction is detected. Right, titration of LDAO detergent with **Cy3-3**. Measurements were performed in triplicates; s.d. around the mean is shown. No interaction is detected in the range up to 6% LDAO. The binding experiments between peptide **Cy3-3** and BamA- β were carried out at 0.6% LDAO concentration. **b**, Left, SDS-PAGE gel of purified OmpX and LamB in LDAO and octyl glucoside (OG) micelles, respectively. Samples in lanes 2 and 4 were boiled, and samples in lanes 1 and 3 were not boiled, before the electrophoresis run. The resulting difference in migration indicates the presence of folded protein in the unboiled samples. For gel source data, see Supplementary Fig. 1. Middle, titration of

LamB and OmpX to **Cy3-3** peptide as monitored by fluorescence anisotropy. No interaction was observed compared to BamA- β . Error bars are s.d. around the mean from triplicate measurements. Right, titration of BamA- β to **Cy3-9** (scrambled) peptide as monitored by fluorescence anisotropy. No interaction was observed compared to **Cy3-3**. Error bars are s.d. around the mean from triplicate measurements. **c**, Binding of **Cy3-3** to BamA- β (left) and to BamA(POTRA1-5) (right) by microscale thermophoresis. Measurements were performed in triplicate with s.d. around the mean shown. For thermophoresis studies of **Cy3-3** with BamA(POTRA1-5), a constant concentration of **Cy3-3** (10 nM) was titrated with BamA(POTRA1-5) in 20 mM HEPES buffer with 150 mM NaCl, pH 7.5, at room temperature, from 107.5 mM to about 3.2 nM. No thermophoresis signal was observed (right).



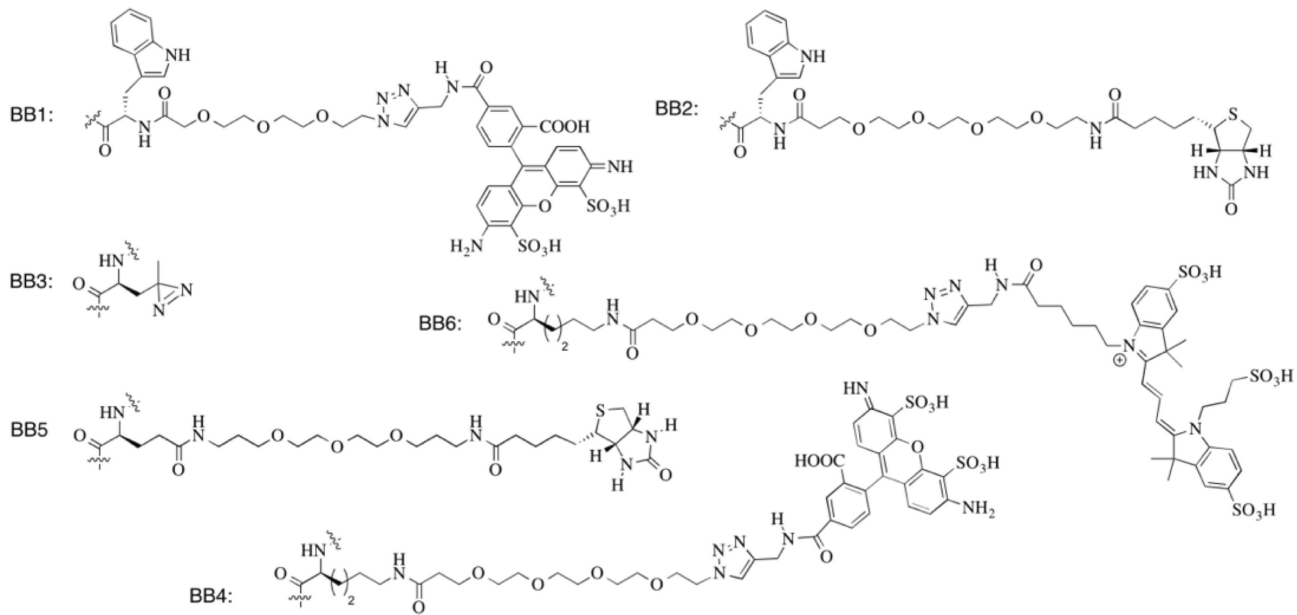
Extended Data Fig. 8 | Interaction-site mapping of BamA^{ext} with peptide 3, peptide 8 and peptide 9. **a**, Interaction-site mapping of BamA^{ext} with peptide 3 ($n=1$ experiment). Chemical shift perturbations of amide moieties plotted against the BamA residue number upon addition of 600 μM peptide 3 to 300 μM of BamA^{ext}. A threshold of 0.07 ppm is indicated by a red dashed line. Three residues, at which the signal goes into intermediate exchange upon peptide titration, are indicated by red lines and their $\Delta\delta(\text{HN})$ was arbitrarily set to 0.1 ppm for visualization. **b**, Chemical shift perturbations of amide moieties plotted against the BamA residue number upon addition of 500 μM peptide 8

to 250 μM of BamA^{ext} ($n=1$ experiment). A threshold of 0.1 ppm is indicated by a red dashed line. Six residues, at which the signal goes into intermediate exchange upon peptide titration, are indicated by red lines and their $\Delta\delta(\text{HN})$ was arbitrarily set to 0.2 ppm for visualization. **c**, Interaction-site mapping of BamA^{ext} with peptide 9 ($n=1$ experiment). Chemical shift perturbations of amide moieties plotted against the BamA residue number upon addition of 700 μM peptide 9 to 350 μM of BamA^{ext}. No statistical tests were done for data shown in this figure.



Extended Data Fig. 9 | Chimeric-antibiotic binding to the BamA β -barrel. The interaction of **8** with the BamA β -barrel domain was characterized using NMR spectroscopy ($n=1$ experiment). **a**, Two-dimensional ^{15}N , ^1H TROSY spectra of 250 μM $[U\text{-}^2\text{H}, ^{15}\text{N}]$ BamA^{ext} (red) titrated with 0.5 (yellow), 1 (magenta) or 2 (blue) stoichiometric equivalents of peptide **8**. **b**, Close-up views of selected residues from the titration in **a**, and from a corresponding titration with BamA- β . **c**, Representation of the interactions of peptide **8** on the BamA β -barrel structure

viewed from the top and from the side of the barrel. Labeled residues have substantial chemical shift perturbations or intensity changes upon peptide binding (crystal structure from PDB 6FSU). **d**, Overlays of two-dimensional ^{15}N , ^1H TROSY spectra of titrations points of BamA^{ext} + scrambled peptide **9**. BamA^{ext} 350 μM (apo), red; +0.5 equivalent of **9**, yellow; +1 equivalent of **9**, pink; and +2 equivalent of **9**, blue.



Extended Data Fig. 10 | Building blocks (BB1–BB6) used in this study. The structures of building blocks BB1 to BB6 used to produce the peptides listed in Extended Data Table 1. For methods of synthesis, see Supplementary Information.

Extended Data Table 1 | Peptides used in this study

peptide	Position																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	Thr	Trp	Ile	Dab	Orn	⁰ Dab	Dab	Trp	Dab	Dab	Ala	Ser	⁰ Pro	Pro										
2	Leu	Ser	Tyr	Dab	Orn	⁰ Dab	Dab	Trp	Dab	tBuGly	Ala	Ser	⁰ Pro	Pro										
3	Val	Thr	Tyr	Dab	Glu*	⁰ Dab	Hse	Trp	Hse	tBuGly	Ala	Ser	⁰ Ala	Pro	Dab	Thr	Dab	Dab*	Dab	⁰ Leu	Leu	Dab	Dab	Thr
fl-3	Val	BB4	Tyr	Dab	Glu*	⁰ Dab	Hse	Trp	Hse	tBuGly	Ala	Ser	⁰ Ala	Pro	Dab	Thr	Dab	Dab*	Dab	⁰ Leu	Leu	Dab	Dab	Thr
Cy3-3	Val	BB6	Tyr	Dab	Glu*	⁰ Dab	Hse	Trp	Hse	tBuGly	Ala	Ser	⁰ Ala	Pro	Dab	Thr	Dab	Dab*	Dab	⁰ Leu	Leu	Dab	Dab	Thr
PAL-3	Val	BB5	Tyr	Dab	Glu*	⁰ Dab	Hse	Trp	Hse	BB3	Ala	Ser	⁰ Ala	Pro	Dab	Thr	Dab	Dab*	Dab	⁰ Leu	Leu	Dab	Dab	Thr
4	Leu	Ser	Tyr	Cys	Gly			Trp	Cys	Val	Ala	Ser	⁰ Pro	Pro	Dab	Thr	Dab	Dab*	Dab	⁰ Leu	Leu	Dab	Dab	Thr
e-4	⁰ Leu	⁰ Ser	⁰ Tyr	⁰ Cys	Gly			⁰ Tryp	⁰ Cys	⁰ Val	⁰ Ala	⁰ Ser	Pro	⁰ Pro	⁰ Dab	⁰ Thr	⁰ Dab	⁰ Dab*	⁰ Dab	Leu	⁰ Leu	⁰ Dab	⁰ Dab	⁰ Thr
fl-4	Leu	Ser	Tyr	Cys	Gly			BB1	Cys	Val	Ala	Ser	⁰ Pro	Pro	Dab	Thr	Dab	Dab*	Dab	⁰ Leu	Leu	Dab	Dab	Thr
PAL-4	Leu	Ser	BB3	Cys	Gly			BB2	Cys	Val	Ala	Ser	⁰ Pro	Pro	Dab	Thr	Dab	Dab*	Dab	⁰ Leu	Leu	Dab	Dab	Thr
5	Leu	Ser	Tyr	Cys	Gly			Trp	Cys	Val	Ala	Ser	⁰ Pro	Pro	DabNH ₂									
6	Colistin nonapeptide														Dab	Thr	Dab	Dab*	Dab	⁰ Leu	Leu	Dab	Dab	Thr
7	AcVal	Cys	Tyr	Dab	Glu*	⁰ Dab	Hse	Trp	Hse	Val	Cys	SerNH ₂			Dab	Thr	Dab	Dab*	Dab	⁰ Leu	Leu	Dab	Dab	Thr
PAL-7	AcVal	Cys	Tyr	BB5	Glu*	⁰ Dab	Hse	Trp	Ser	BB3	Cys	SerNH ₂			Dab	Thr	Dab	Dab*	Dab	⁰ Leu	Leu	Dab	Dab	Thr
8	AcVal	Cys	Tyr	T	Glu*	⁰ Dab	U	Trp	X	Val	Cys	Y			Dab	Thr	Z	Dab*	Dab	⁰ Leu	Leu	Dab	Dab	Thr
e-8	Ac ⁰ Val	⁰ Cys	⁰ Tyr	⁰ T	⁰ Glu*	Dab	⁰ U	⁰ Trp	⁰ X	⁰ Val	⁰ Cys	⁰ Y			⁰ Dab	⁰ Thr	⁰ Z	⁰ Dab*	⁰ Dab	Leu	⁰ Leu	⁰ Dab	⁰ Dab	⁰ Thr
9	Val	Tyr	Thr	tBuGly	Glu*	⁰ Dab	Hse	Trp	Hse	Dab	Pro	Ser	⁰ Ala	Ala	Dab	Thr	Dab	Dab*	Leu	Dab	Dab	⁰ Leu	Dab	Thr
Glu*: side chain C cross-linked with N Dab ¹⁵ ; Dab*: side chain N cross-linked with C Thr ²⁴																								

The amino acid residues or other building blocks (BB1 to BB6) (Extended Data Fig. 10) at each position indicated in the structures in Fig. 1. Glu* and Dab* indicate peptide linkage through the side chain of Glu and Dab, respectively (Fig. 1). Dap, L-2,3-diaminopropionic acid. The letters T, U, X, Y and Z indicate variable positions. For methods of synthesis, see Supplementary Information.

A new antibiotic selectively kills Gram-negative pathogens

<https://doi.org/10.1038/s41586-019-1791-1>

Received: 3 April 2019

Accepted: 8 November 2019

Published online: 20 November 2019

Yu Imai^{1,15}, Kirsten J. Meyer^{1,15}, Akira Inishi¹, Quentin Favre-Godal¹, Robert Green¹, Sylvie Manuse¹, Mariaelena Caboni¹, Miho Mori¹, Samantha Niles¹, Meghan Ghiglieri¹, Chandrashekhar Honrao², Xiaoyu Ma², Jason J. Guo^{2,3}, Alexandros Makriyannis², Luis Linares-Otoya⁴, Nils Böhlinger⁴, Zerlina G. Wuisan⁴, Hundeeep Kaur⁵, Runrun Wu^{6,7}, André Mateus⁸, Athanasios Typas⁸, Mikhail M. Savitski⁸, Josh L. Espinoza^{9,10}, Aubrie O'Rourke^{9,10}, Karen E. Nelson^{9,10,11,12}, Sebastian Hiller⁵, Nicholas Noinaj^{6,7}, Till F. Schäberle^{4,13,14}, Anthony D'Onofrio¹ & Kim Lewis^{1*}

The current need for novel antibiotics is especially acute for drug-resistant Gram-negative pathogens^{1,2}. These microorganisms have a highly restrictive permeability barrier, which limits the penetration of most compounds^{3,4}. As a result, the last class of antibiotics that acted against Gram-negative bacteria was developed in the 1960s². We reason that useful compounds can be found in bacteria that share similar requirements for antibiotics with humans, and focus on *Photorhabdus* symbionts of entomopathogenic nematode microbiomes. Here we report a new antibiotic that we name darobactin, which was obtained using a screen of *Photorhabdus* isolates. Darobactin is coded by a silent operon with little production under laboratory conditions, and is ribosomally synthesized. Darobactin has an unusual structure with two fused rings that form post-translationally. The compound is active against important Gram-negative pathogens both in vitro and in animal models of infection. Mutants that are resistant to darobactin map to BamA, an essential chaperone and translocator that folds outer membrane proteins. Our study suggests that bacterial symbionts of animals contain antibiotics that are particularly suitable for development into therapeutics.

It is difficult to find compounds that target Gram-negative bacteria^{1,2}. This problem is largely responsible for the current antimicrobial resistance crisis. Pathogens such as *Escherichia coli*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa* and *Acinetobacter baumannii* have acquired resistance to most—and in some cases to all—antibiotics that are currently available in the clinic. The WHO (World Health Organization) has recently classified these drug-resistant pathogens as a critical priority for global human health⁵.

Gram-negative bacteria evolved an outer membrane to protect themselves from unwanted compounds^{3,4}. Only a small number of antibiotics can penetrate this barrier, and are active against Gram-negative bacteria.

Most of these compounds are natural products that are made by soil microorganisms, and mainly by Actinomycetes—aminoglycosides, tetracyclines and β -lactams. The last class of antibiotics to act against Gram-negative bacteria, the synthetic fluoroquinolones, were introduced half a century ago. Since then, discovery of

new antibiotics has largely been limited to narrow-spectrum compounds^{2,6}.

We reasoned that useful compounds will be present in microorganisms that have the same requirements for antibiotics as humans. The nematode symbionts *Photorhabdus* and *Xenorhabdus* seem to represent such a group of microorganisms. These nematophilic bacteria are members of the gut microbiome of nematodes and are closely related to other Enterobacteriaceae, such as *E. coli*. Nematodes invade insect larvae and release their symbionts. Nematophilic bacteria first produce neurotoxins to immobilize the prey, and then release various antimicrobials to fend off invading environmental microorganisms^{7,8}. However, the most-direct competitors probably do not come from the environment, but from other members of the nematode gut microbiome. Notably, Gram-negative bacteria that are common opportunistic pathogens of humans are abundant in the microbiome of entomopathogenic nematodes⁹. The antimicrobial compounds produced by nematophilic bacteria must be non-toxic to the nematode, and be able to

¹Antimicrobial Discovery Center, Department of Biology, Northeastern University, Boston, MA, USA. ²Center for Drug Discovery, Department of Pharmaceutical Sciences, Northeastern University, Boston, MA, USA. ³Barnett Institute for Chemical and Biological Analysis, Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, USA. ⁴Institute for Insect Biotechnology, Justus-Liebig-University of Giessen, Giessen, Germany. ⁵Biozentrum, University of Basel, Basel, Switzerland. ⁶Purdue Institute of Inflammation, Immunology and Infectious Disease, Purdue University, West Lafayette, IN, USA. ⁷Markey Center for Structural Biology, Department of Biological Sciences, Purdue University, West Lafayette, IN, USA. ⁸Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. ⁹Department of Human Biology, J. Craig Venter Institute, La Jolla, CA, USA. ¹⁰Department of Genomic Medicine, J. Craig Venter Institute, La Jolla, CA, USA. ¹¹Department of Human Biology, J. Craig Venter Institute, Rockville, MD, USA. ¹²Department of Genomic Medicine, J. Craig Venter Institute, Rockville, MD, USA. ¹³Department of Bioresources, Fraunhofer Institute for Molecular Biology and Applied Ecology, Giessen, Germany. ¹⁴German Center for Infection Research (DZIF), Partner Site Giessen-Marburg-Langen, Giessen, Germany. ¹⁵These authors contributed equally: Yu Imai, Kirsten J. Meyer. *e-mail: k.lewis@neu.edu

spread well through the insect larvae. This suggests the production of antimicrobials with low toxicity and good pharmacokinetics that are active against Gram-negative pathogens.

Identification of darobactin

We screened a small set of *Photorhabdus* and *Xenorhabdus* strains, including a total of 67 isolates from 28 species (Extended Data Table 1), against *E. coli*. Usually, antibiotic-producing bacteria are spotted onto a nutrient agar plate overlaid with a target microorganism. Most of the tested bacteria did not produce zones of inhibition, and we reasoned that this may be due to poor expression of 'silent' biosynthetic gene clusters (BGCs) in vitro. We therefore prepared concentrated extracts from the bacterial cultures and spotted them on overlay plates. A concentrated extract from *Photorhabdus kharii* HGB1456 produced a small zone of *E. coli* growth inhibition on a Petri dish, whereas spotting a colony of *P. kharii* HGB1456 had no effect on the growth of *E. coli* (Fig. 1a). Bioassay-guided isolation of the extract by high-performance liquid chromatography produced an active fraction (Extended Data Fig. 1a). High-resolution electrospray ionization–mass spectrometry analysis identified a compound with a molecular mass of 966.41047, which is consistent with a molecular formula of $C_{47}H_{56}O_{12}N_{11}^+$ (calculated $[M+H]^+ = 966.41044$). This mass did not have a match in Antibase, suggesting the presence of a novel compound. Mass-spectrometry fragmentation and nuclear magnetic resonance (NMR) studies (Extended Data Fig. 1b–h) led to the identification of the structure of the active compound, which we named darobactin (Fig. 1b). Darobactin is a modified heptapeptide with an amino acid sequence of $W^1-N^2-W^3-S^4-K^5-S^6-F^7$. NMR studies revealed two unusual macrocycle crosslinks in darobactin: a previously undescribed aromatic–aliphatic ether link between the C7 indole of W^1 and the β -carbon of W^3 , and a carbon–carbon bond between the C6 indole of W^3 and the β -carbon of K^5 . The Trp–Lys bond is made between two unactivated carbons, which has not been described previously for an antibiotic. We next sequenced the genome of *P. kharii* HGB1456 (GenBank accession number WHZZ000000000) and searched for BGCs that encode non-ribosomal peptide synthetases. There were 10 non-ribosomal peptide synthetases in the genome, but none of them could be predicted to form the darobactin peptide. Next, we directly compared the sequence of this seven amino acid peptide with the genome of *P. kharii* and found a perfect match near the C terminus of an open-reading frame that encodes a peptide with a length of 58 amino acids. The ribosomal synthesis of darobactin suggests that the amino acid backbone is in the L-configuration. The macrocycle crosslinks generate two chiral centres at the β -carbons of W^3 and K^5 , which have *R* and *S* configurations, respectively, based on nuclear Overhauser effect correlations and molecular modelling (Extended Data Fig. 2).

The putative operon that encodes darobactin (Fig. 1c and Extended Data Fig. 3) is typical of ribosomally synthesized and post-translationally modified peptide genes (RiPPs) that encode a variety of ribosomally produced natural products, including nisin—a food preservative—and the antibiotic thiostrepton. This *dar* operon consists of a propeptide encoded by *darA*, a small *relE*-type open-reading frame, *darBCD*—which encodes an ABC-type trans-envelope exporter (*darB* and *darD* make up the transporter itself, and *darC* encodes a membrane fusion protein)—and *darE*, which encodes a radical S-adenosyl methionine (SAM) enzyme. Enzymes of the radical SAM class catalyse free-radical-based reactions that can link unactivated carbons¹⁰. This explains the formation of the Trp–Lys C–C bond in darobactin. Such a Trp–Lys C–C bond was recently reported in a peptide pheromone, streptide, from *Streptococcus thermophilus*¹¹. There is little overall homology between the two enzymes, but DarE contains the SAM and SPASM domains that are characteristic of this group. The operon does not contain a separate enzyme for generating the ether bond in the first ring. RiPP operons often encode a protease that cleaves out the active peptide; however, this was not present in the *dar* operon. Hence,

Table 1 | MIC and cytotoxicity of darobactin against pathogens, intestinal gut bacteria and human cell lines

Organism and genotype	Concentration ($\mu\text{g ml}^{-1}$)	
	Darobactin	Ampicillin
Pathogenic bacteria (MIC)		
<i>Pseudomonas aeruginosa</i> PAO1	2	>128
<i>Pseudomonas aeruginosa</i> pmrB 523C>T	2	>128
<i>Pseudomonas aeruginosa</i> JMI 1045324	16	ND
<i>Shigella sonnei</i> ATCC 25931 ^a	2	4
<i>Klebsiella pneumoniae</i> ATCC 700603	2	128
<i>Klebsiella pneumoniae</i> ESK JMI 1052654	2	>128
<i>Klebsiella pneumoniae</i> ATCC 700603 (SHV-18)	4	>128
<i>Klebsiella pneumoniae</i> ATCC BAA-1705 (KPC)	4	>128
<i>Escherichia coli</i> ATCC 25922	2	8
<i>Escherichia coli</i> AR350 (<i>mcr-1</i>)	2	>128
<i>Escherichia coli</i> ESK JMI 1043856	2	>128
<i>Escherichia coli</i> ATCC BAA-2340 (KPC)	2	>128
<i>Escherichia coli</i> MG1655 +10% serum	2	4
<i>Escherichia coli</i> MG1655	4	4
<i>Salmonella</i> Typhimurium LT2 ATCC 19585 ^a	4	2
<i>Moraxella catarrhalis</i> ATCC 25238 ^a	8	<0.25
<i>Acinetobacter baumannii</i> ATCC 17978	8	64
<i>Enterobacter cloacae</i> ATCC 13047 ^a	32	>128
<i>Proteus mirabilis</i> KLE 2600 ^{a,b}	64	>128
<i>Staphylococcus aureus</i> HG003	>128	0.5
<i>Clostridium bifermentans</i> KLE 2329 ^{a,b}	>128	1
<i>Mycobacterium tuberculosis</i> mc ² 6020	>128	16
Symbiotic gut bacteria (MIC)		
<i>Bifidobacterium longum</i> ATCC BAA-999 ^a	>128	0.25
<i>Bacteroides fragilis</i> ATCC 25285 ^a	>128	128
<i>Bacteroides xylanisolvens</i> KLE 2253 ^{a,b}	>128	1
<i>Bacteroides dorei</i> KLE 2422 ^{a,b}	>128	1
<i>Bacteroides caccae</i> KLE 2423 ^{a,b}	>128	2
<i>Bacteroides vulgatus</i> KLE 2303 ^{a,b}	>128	2
<i>Bacteroides nordii</i> KLE 2369 ^{a,b}	>128	4
<i>Lactobacillus reuteri</i> ATCC 23272 ^a	>128	1
<i>Enterococcus faecalis</i> KLE 2341 ^{a,b}	>128	4
<i>Faecalibacterium prausnitzii</i> KLE 2243 ^{a,b}	>128	64
<i>Haemophilus parainfluenzae</i> KLE 2367 ^{a,b}	>128	128
<i>Stenotrophomonas maltophilia</i> KLE 11416 ^{a,b}	>128	>128
Human cell line (IC₅₀)		
HepG2	>128	>128
FaDu	>128	>128
HEK293	>128	>128

ND, no data. ESK, extended spectrum β -lactamase.

^aCultivated under anaerobic conditions.

^bHuman stool isolate, K.L. laboratory collection.

generic proteolysis may be involved in the maturation of the propeptide. To link the putative BGC with production of darobactin, we generated a markerless knockout mutant in which the complete *darABCDE* operon was deleted from *P. kharii* DSM3369 by double crossover. Darobactin production was abolished in the resulting mutant strain (Extended Data Fig. 4a, c, d). Notably, darobactin was produced heterologously from the *dar* operon cloned into *E. coli* (Extended Data Fig. 4b, d). This shows that the *dar* operon is sufficient for making darobactin. Furthermore, it appears that the DarE radical SAM enzyme

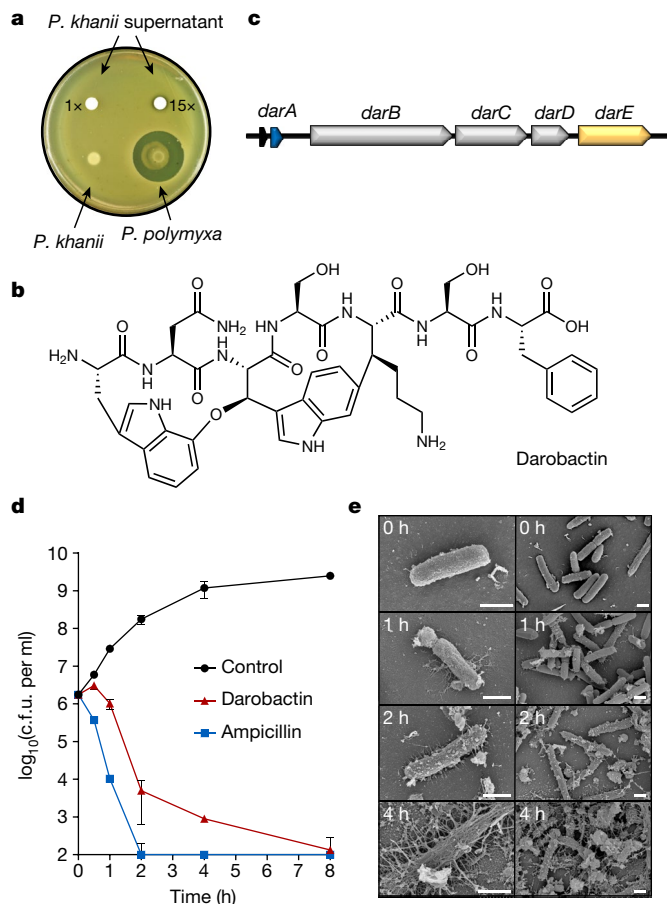


Fig. 1 | Darobactin produced by a silent operon of *P. khanii* is a bactericidal antibiotic. **a**, *P. khanii* was grown in liquid culture, after which concentrated culture supernatants were tested for inhibition of *E. coli* MG1655. The concentrated supernatant of *P. khanii* produced a zone of inhibition on an *E. coli* lawn, whereas unconcentrated supernatant or a colony overlay did not. *Paenibacillus polymyxa* produces polymyxin and serves as a positive control. **b**, Darobactin structure. **c**, The BGC consists of the structural gene *darA* (coloured in blue), *darBCD* (transporter encoding genes; grey) and *darE* (encoding a radical SAM enzyme; orange). In addition, a *relE*-like gene (black) open-reading frame is co-located with the BGC at different positions in different species. **d**, Time-dependent killing of *E. coli* MG1655 by darobactin. An exponential culture of *E. coli* MG1655 was challenged with 16× MIC antibiotics. *n* = 3 biologically independent samples. Data are mean ± s.d. c.f.u., colony-forming units. **e**, Scanning electron microscopy analysis of *E. coli* MG1655 treated with 16× MIC darobactin. Scale bars, 1 μm.

catalyses the formation of both the Trp–Lys C–C bond and the C–O–C Trp–Trp ether bond. The chemistries of these two reactions are substantially different, and the mechanism of DarE catalysis requires a separate investigation.

We find that the *dar* operon is common in *Photorhabdus*, and detected it in 15 different species for which the genome sequence is available (Extended Data Fig. 4e). The *dar* operon was absent only in *Photorhabdus bodei*. Synteny of the genomes that contain the *dar* locus with that of *P. bodei* helped to determine the boundaries of the operon (Extended Data Fig. 3a, c). We also tested the production of darobactin in several different *Photorhabdus* species, and found that it is the highest in a strain of *P. khanii* DSM 3369. We switched to this strain for the isolation of darobactin; however, even in this isolate, the production of darobactin is low (3 mg l⁻¹), only twofold higher than in *P. khanii* HGB1456, and requires unusually long fermentation (10–14 days). This probably explains why darobactin has been overlooked in screens for antibiotics.

We then expanded the search for *dar*-type operons in databases of bacterial genome sequences (NCBI), using the propeptide and the *dar*-encoding peptide as queries. The two searches identified homologues of the *dar* operon that appear to encode four darobactin analogues. We therefore propose the name darobactin A for the first compound, and darobactin B–E for the predicted analogues of this class of antibiotics. In *Photorhabdus australis* and *Photorhabdus asymbiotica*, the sequence data suggest the presence of darobactin B, which contains two amino acid changes in the N terminus (SKSF to TKRF). In multiple *Yersinia* species, the second amino acid (N to S), the fifth amino acid (K to R), or both, are modified. We named these analogues darobactin C, D and E, respectively (Extended Data Fig. 4e, f). Notably, the sequence of darobactin C is present in *Yersinia pestis*, the causative agent of the plague, and in *Yersinia frederiksenii* from the human gut microbiome. Darobactin A is the most common, and a corresponding propeptide sequence is present in six sequenced *Photorhabdus* species, seven *Yersinia* species, *Vibrio crassostreae* and *Pseudalteromonas luteoviolacea*, all of which are γ-proteobacteria. All species that contain *dar* operons are associated with animals. Apparently, combinatorial reshuffling of the *dar* operon produced a family of genes, and the five analogues were selected over the course of evolution from a total of 1.28×10^9 (20⁷) sequences. The GC content of the *dar* operon is 32%, significantly lower than the rest of the genomes of *P. khanii* and other γ-proteobacteria, which have a GC content of 45%. This suggests that the operon was horizontally acquired from a microorganism in which darobactin evolved. Although the nature of this microorganism is unknown, it is not an actinomycete—their genomes have a characteristically high GC content (>55%)¹².

Identifying the target

Darobactin had reasonable activity against a range of Gram-negative bacteria, with a minimum inhibitory concentration (MIC) of 2 μg ml⁻¹ against important drug-resistant pathogens, *E. coli* and *K. pneumoniae*, including polymyxin-resistant, extended spectrum β-lactamase and carbapenem-resistant clinical isolates (Table 1 and Supplementary Table 1). The compound is bactericidal (Fig. 1d), with a minimal bactericidal concentration of 8 μg ml⁻¹ against *E. coli*. There was little activity against Gram-positive bacteria. Notably, the compound was also largely inactive against gut commensals, including *Bacteroides*, the main group of Gram-negative symbionts¹³. Disrupting the microbiome by antibiotics, especially early in life, is a major concern, given the important role of symbiotic bacteria in many aspects of human health, such as shaping the immune system during development¹⁴.

Darobactin is a large, 965 Da, molecule, whereas the cut-off for compounds to permeate the outer membrane¹⁵ is around 600 Da. We therefore considered that darobactin, similarly to polymyxin, might target the lipopolysaccharides of the outer membrane. Adding purified lipopolysaccharides to a culture of *E. coli* protected cells from polymyxin, but had no effect on darobactin activity (Extended Data Fig. 5a). Addition of darobactin to *E. coli* caused blebbing of the membrane, and eventual swelling and lysis of cells (Fig. 1e, Extended Data Fig. 6 and Supplementary Video 1). Transcriptome analysis revealed that darobactin rapidly (in 15–30 min) induced the sigma E and Rcs envelope stress responses, and more broadly activated genes from all five envelope stress pathways (Extended Data Fig. 7 and Supplementary Discussion). To identify the target of darobactin, we performed a ligand-protection thermal proteome analysis. This, however, did not reveal a particular protein the denaturation of which was protected by darobactin. At the same time, the proteome showed that the abundance of periplasmic chaperones Spy and DegP was markedly increased, and that the abundance of outer membrane proteins, especially NanC, LamB and OmpF, was decreased (at least in part due to a decrease in the respective transcripts) in response to darobactin treatment (Extended Data Fig. 8, Supplementary Table 2 and Supplementary Discussion). Microscopy, transcriptome and proteome

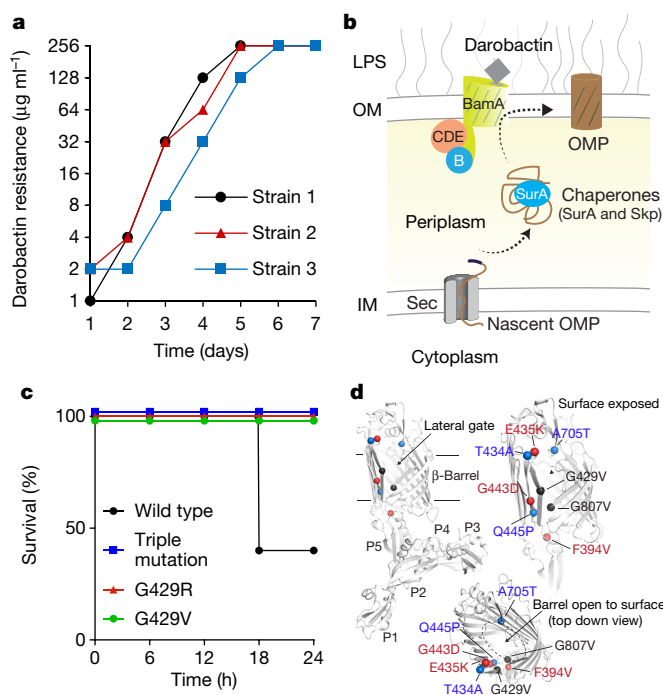


Fig. 2 | Multiple mutations in *bamA* confer darobactin resistance.

a, Darobactin-resistant mutants were generated by daily serial passage of *E. coli* MG1655 at sub-MIC concentrations of darobactin, leading to a steady shift in the darobactin concentration that permits *E. coli* MG1655 growth. This experiment was performed in three biologically independent samples. The three mutants obtained contained 2–3 mutations in *bamA*. **b**, Schematic of the Bam complex³⁵. IM, inner membrane; OM, outer membrane. **c**, Mice were injected with 10^7 c.f.u. of *E. coli* ATCC 25922 wild-type or darobactin-resistant strains. Resistant strains carried either; the triple mutations evolved in strain 3 (**a**), or single spontaneous mutations of G429 to R or V. $n = 5$ per group. Mice were monitored for survival. **d**, Darobactin-resistance mutations (coloured spheres) mapped to the BamA protein structure (grey) shown as a cartoon with the β -barrel domain and the individual polypeptide-transport-associated domains indicated.

analyses indicate a defect in the cell envelope. We next sought to obtain mutants that were resistant to darobactin to identify its target. Plating *E. coli* on solid medium that contained darobactin at $4 \times \text{MIC}$ produced resistant mutants with a frequency of 8×10^{-9} . To obtain mutants that were resistant to higher levels of the compound, we performed an evolutionary experiment in liquid medium¹⁶ (Fig. 2a). Repeatedly re-inoculating a culture into medium with progressively increasing levels of the antibiotic produced mutants with a high resistance to darobactin, which had MICs greater than $128 \mu\text{g ml}^{-1}$ (Fig. 2a). Sequencing the mutants showed that in all three strains of *E. coli*, there were 2–3 mutations in the gene that encodes BamA, an essential outer membrane protein¹⁷ (Fig. 2b). After transferring the three *bamA* mutations from the resistant strain 3 to a clean *E. coli* background by allelic replacement, we confirmed that they are solely responsible for darobactin resistance (MIC of $128 \mu\text{g ml}^{-1}$). The ability to generate mutants that are resistant to high levels of the compound suggests a lack of off-target activity. To sustain an infection in the presence of an antibiotic, the pathogen should be both resistant and virulent. We therefore tested whether darobactin-resistant mutants retained virulence. Injecting mice with 10^7 cells of *E. coli* ATCC 25922 caused 60% mortality within 24 h. By contrast, there was no death at 24 h when the animals were inoculated with *E. coli* carrying either single or triple mutations in *bamA* (Fig. 2c and Extended Data Fig. 5b). *E. coli* virulence is thus strongly compromised by *bamA* mutations that confer resistance to darobactin.

BamA is the central component of the BamABCDE complex¹⁷ (Fig. 2b). One proposed mechanism for BAM is that nascent outer membrane proteins are inserted from the periplasm into the outer membrane by the central component BamA, which serves to catalyse both folding and insertion. BamA is not an enzyme, and its β -barrel structure does not obviously lend itself to inhibition by small molecules. BamA is targeted by large lectin-like bacteriocins, LlpA¹⁸, and a group from Genentech developed an antibody that inhibits this protein in *E. coli*¹⁹. In a recent study, a small molecule synthetic compound MRL-494 was reported to act against BamA, without the need to penetrate the outer membrane²⁰. MRL-494 is active against *E. coli* and *K. pneumoniae* with an MIC of 15 and $62 \mu\text{g ml}^{-1}$, respectively, whereas it acts against Gram-positive bacteria by disrupting their cytoplasmic membrane.

We observed direct inhibition of BAM by darobactin using an in vitro protein refolding assay. The isolated BAM complex was integrated into lipid nanodiscs, and its ability to fold the protease OmpT was measured (Extended Data Fig. 5c). Darobactin inhibited BAM-dependent folding of OmpT with an apparent half-maximum inhibitory concentration (IC_{50}) of $0.68\text{--}1 \mu\text{M}$ (Fig. 3a and Extended Data Fig. 5d), consistent with the MIC of $1.9 \mu\text{M}$ in *E. coli*. Darobactin had no effect on OmpT activity in the absence of BAM (Extended Data Fig. 5e), and a linear peptide with the same sequence as darobactin had no inhibitory activity on BAM (Extended Data Fig. 5f). We next tested darobactin-resistant mutants in the same assay. The IC_{50} of mutant 1a was increased markedly, to $120 \mu\text{M}$. In mutants 2 and 3, the IC_{50} was unchanged, but the folding activity was strongly decreased (Fig. 3a). The mechanism by which mutants 2 and 3 confer resistance is unclear and will require additional study.

Using isothermal titration calorimetry experiments, we also observed that darobactin directly and specifically interacts with BamA of BAM with a measured dissociation constant (K_d) of $1.2 \mu\text{M}$, with no binding observed for the linear peptide (Fig. 3b and Extended Data Fig. 5g, h).

To characterize the interaction of BamA with darobactin at the atomic level, we performed a high-resolution NMR study. Stepwise titration of the unlabelled darobactin with the [^{15}N , ^2H]-labelled BamA β -barrel (BamA- β) was carried out and monitored by solution NMR spectroscopy. Upon addition of 0.5-molar-equivalent darobactin, significant changes were observed in the NMR spectrum of BamA- β , which became more prominent at 1 molar equivalent (Extended Data Fig. 5i and Supplementary Data 1). By contrast, a linear scrambled darobactin peptide had no effect on the NMR spectrum (Extended Data Fig. 5j and Supplementary Data 1). We have previously shown that BamA- β exists as an interchanging two-state ensemble of a gate-closed and a gate-opened conformation and that each of these two conformations can be stabilized by a conformation-specific nanobody, nanoF7 for the gate-closed and nanoE6 for the gate-opened structure^{21,22}. Notably, we found that darobactin stabilized one of these two conformations (Fig. 3c, d). The darobactin-stabilized conformation resembles for most residues the closed-gate conformation, as shown by the high similarity of NMR spectral positions of BamA- β and nanoF7 and BamA- β and darobactin, whereas the NMR position of BamA- β and nanoE6 was clearly different (Fig. 3c, d). These findings strongly suggest that darobactin stabilizes a closed lateral gate upon binding to BamA, preventing the exit of substrates into the outer membrane. Notably, most mutations that confer resistance to darobactin are located at the lateral gate of BamA (Fig. 2d). Taken together, these findings are consistent with darobactin inhibiting BamA and disrupting the formation of a functional outer membrane. Future studies will determine the mechanism by which darobactin kills bacterial cells by acting against this target.

Animal efficacy

Given the attractive mode of action and lack of cytotoxicity (Table 1), we next examined the efficacy of darobactin in mouse models of infection. Single-dose pharmacokinetic analysis shows that darobactin achieves good exposure, with an intraperitoneal injection of 50 mg kg^{-1} leading

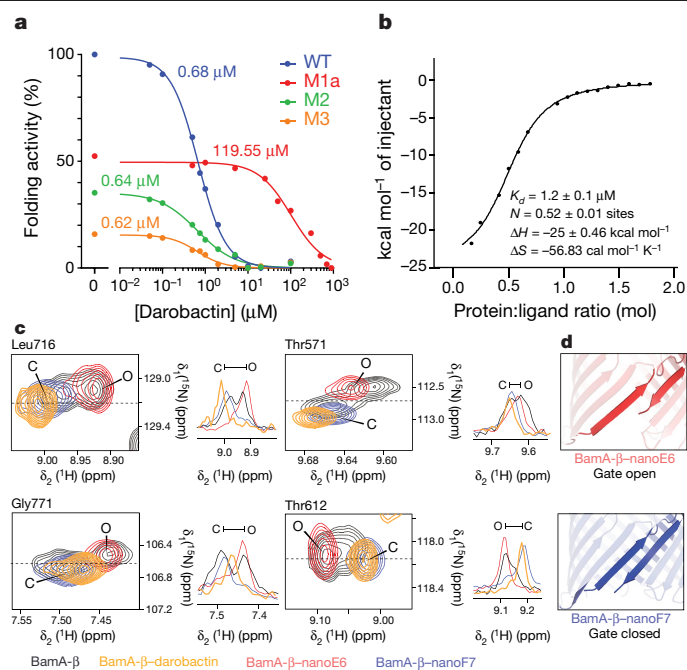


Fig. 3 | Darobactin inhibits BAM activity, and binds to and induces selection of the closed-gate conformation of BamA-β. **a**, The assay shown in Extended Data Fig. 5c was used to measure BAM activity in wild-type (WT) and darobactin-resistant mutants, in the presence of increasing concentrations of darobactin. The IC_{50} values are indicated for each mutant. The 95% confidence intervals for the IC_{50} are: wild type 0.61–0.75 μ M, M1a (G429V, T434A and G807V; Methods) 68–148 μ M, M2 (F394V, E435K and G443D) 0.50–0.83 μ M, M3 (T434A, Q445P and A705T) 0.38–0.94 μ M (GraphPad Prism v.8.2). The experiment was repeated independently at least three times with similar results. **b**, Specific binding of darobactin to BamA/BAM. Plot of isothermal titration calorimetry experiments of wild-type BAM titrated with darobactin. $K_d = 1.2 \mu$ M, $N = 0.52$, $\Delta H = -25 \text{ kcal mol}^{-1}$ and $\Delta S = -56 \text{ cal mol}^{-1} \text{ K}^{-1}$. The experiment was repeated independently twice with similar results. **c**, Two-dimensional magnification and one-dimensional cross-sections from two-dimensional ^{15}N , ^1H -TROSY spectra of BamA-β in lauryldimethylamine-*N*-oxide micelles for four selected amino acid residues, as indicated at the top of each panel. Apo BamA-β (black), equimolar BamA-β-darobactin (orange), BamA-β-nanoF7 (blue) and BamA-β-nanoE6 (red). Resonances that correspond to the open and closed conformation are indicated as O and C, respectively. The experiment was repeated independently twice with similar results. **d**, Conformation of the gate region in crystal structures of BamA-β-nanoE6 and BamA-β-nanoF7 (Protein Data Bank (PDB) 6QGW and 6QGX7, respectively).

to a peak blood level of 94 $\mu\text{g ml}^{-1}$ and a half-life of 1 h (Extended Data Fig. 9a). Notably, the blood levels of the compound were maintained above the MIC of *E. coli* for 8 h, an excellent predictor of efficacy. We also did not notice any toxicity with this dose of darobactin. Next, the efficacy of the compound was examined in a mouse septicemia model. For this, we examined wild-type and polymyxin-resistant *P. aeruginosa* (PAO1 and *pmrB* 523C>T), carbapenemase-producing *K. pneumoniae* (KPC), and wild-type and polymyxin-resistant *E. coli* (ATCC 25922 and AR350 *mcr-1*) (Fig. 4a–c). Carbapenem-resistant *K. pneumoniae* causes 30–40% mortality in the United States and 40–50% in Europe^{23,24}. Polymyxin-resistant *E. coli mcr-1* is of particular concern, as the resistance locus is present on a plasmid and can rapidly spread²⁵.

To initiate septicemia, mice were infected intraperitoneally and 1 h after introducing the pathogens, darobactin was administered. Untreated controls all died within 24 h, but a single dose of darobactin completely protected the animals infected with *E. coli*, *K. pneumoniae* and polymyxin-resistant *P. aeruginosa* (Fig. 4a–c). Darobactin given in three doses of 25 mg kg^{-1} cured two out of three mice infected with

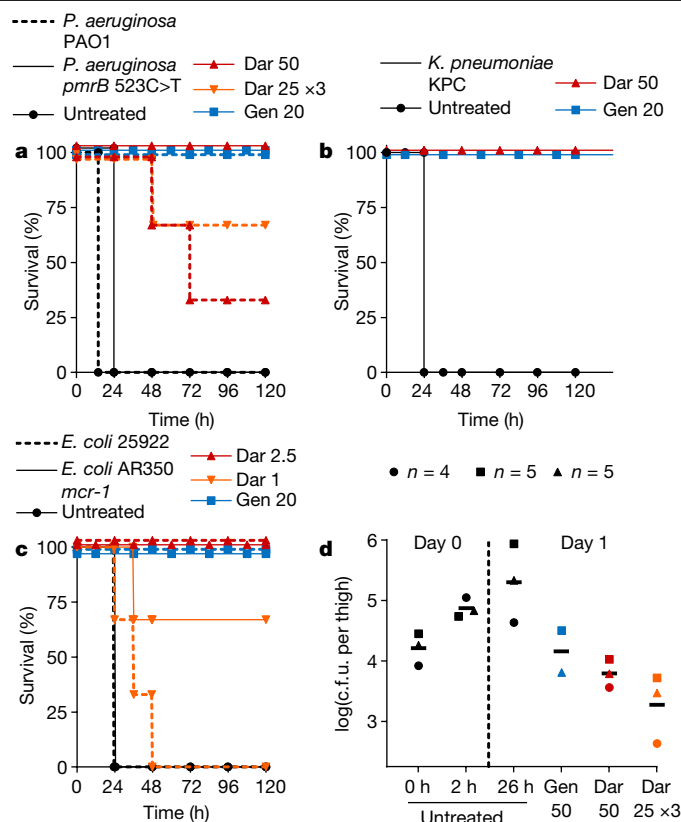


Fig. 4 | Darobactin is efficacious in mouse infection models. **a–c**, Mice were given a lethal inoculum of bacteria (intraperitoneal injection) and antibiotics were administered 1 h later. **a**, Darobactin (Dar) was tested against PAO1 wild-type and *pmrB* 523C>T (resistant to polymyxin) *P. aeruginosa* septicemia, $n = 3$ per group. ‘25 \times 3’ refers to three doses given every 6 h. **b**, Darobactin was tested against carbapenemase-producing *K. pneumoniae* (KPC), $n = 3$ per group. **c**, Determining the minimum curative dose of darobactin against wild-type *E. coli* (ATCC 25922) and the polymyxin-resistant clinical isolate (AR350), $n = 3$ per group. **d**, In a neutropenic thigh model, darobactin was given as a single dose (intraperitoneal injection) at 2 h after infection, or administered three times at 2, 8 and 14 h after infection. The right quadriceps muscle was collected, homogenized, serially diluted and plated for c.f.u. analysis at 26 h. The experiment was repeated three times, symbols represent the average of the group in each experiment ($n = 4$ or 5), lines are the mean of experiments. Gentamicin (Gen) was used as a positive control. All treatments are in mg kg^{-1} .

wild-type *P. aeruginosa* PAO1 (Fig. 4a). Darobactin was then tested in a mouse thigh infection with *E. coli mcr-1*. In this model, animals were made neutropenic with cyclophosphamide treatment, and the ability of the antibiotic to kill the pathogen is tested in the absence of an immune response. Darobactin, given as either a single injection of 50 mg kg^{-1} or as three injections of 25 mg kg^{-1} every 6 h, significantly decreased the pathogen burden at 26 h, and was more efficacious than gentamicin (50 mg kg^{-1}) (Fig. 4d and Extended Data Fig. 9b). These experiments suggest that darobactin is a promising lead compound for developing a therapeutic against Gram-negative pathogens.

Discussion

The number of novel compounds that target Gram-negative bacteria is small, comprising mainly β -lactamase inhibitors—avibactam²⁶, vaborbactam²⁷ and aspergillomarasmine²⁸; arylomycin analogues that target the LepB signal peptidase²⁹ are in development by Genentech³⁰.

An intriguing new discovery platform is in development, based on emerging rules of permeation that determine the properties that are required for compounds to breach the permeability barrier

of Gram-negative bacteria³¹. However, perhaps the most-practical approach is currently to mine untapped groups of microorganisms that may harbour new chemistry. These include uncultured bacteria, from which teixobactin was discovered¹⁶; and several of the most abundant soil taxa—Acidobacteria, Verrucomicrobia, Rokubacteria and Gemmatimonadetes³². Several dozen compounds with antimicrobial properties have been isolated from nematophilic bacteria³³, but these do not hit specific targets or do not appear to have drug-like properties. The currently identified compounds represent a small fraction of what is encoded by the genomes of nematophilic bacteria and are expressed well under laboratory conditions; darobactin is encoded by a silent operon. Nematophilic bacteria split from other Enterobacteriaceae around 370 million years ago³⁴. Since then, they would have acquired, by horizontal transmission from the biosphere, antibiotics that may be of use to us.

Darobactin is indeed a typical example of a compound that is acquired by horizontal transmission of a BGC operon from an unknown microorganism. It acts against an attractive target on the surface of the cell. The BamA chaperone, which is itself an outer membrane β -barrel protein, catalyses folding and insertion of new β -barrel proteins into the outer membrane. Drugs in general, and natural products in particular, normally target enzymes with their well-defined catalytic centres, rather than chaperones. According to our data, darobactin stabilizes the closed lateral gate conformation of BamA, preventing it from opening and inserting its substrates into the membrane. Darobactin is a large molecule, which is probably necessary for its unusual mode of action. The location of the target on the surface resolves the intractable problem of penetration across the permeability barrier of Gram-negative bacteria. There are only two essential proteins exposed on the surface of the outer membrane—BamA and LptD¹⁷. There is little doubt that nature produced more than one type of compounds that acts against these targets.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1791-1>.

- Payne, D. J., Gwynn, M. N., Holmes, D. J. & Pompliano, D. L. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discov.* **6**, 29–40 (2007).
- Lewis, K. Platforms for antibiotic discovery. *Nat. Rev. Drug Discov.* **12**, 371–387 (2013).
- Lomovskaya, O. & Lewis, K. *emr*, an *Escherichia coli* locus for multidrug resistance. *Proc. Natl Acad. Sci. USA* **89**, 8938–8942 (1992).
- Li, X. Z. & Nikaido, H. Efflux-mediated drug resistance in bacteria. *Drugs* **64**, 159–204 (2004).
- Tacconelli, E. et al. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect. Dis.* **18**, 318–327 (2018).
- Brown, E. D. & Wright, G. D. Antibacterial drug discovery in the resistance era. *Nature* **529**, 336–343 (2016).
- Crawford, J. M. & Clardy, J. Bacterial symbionts and natural products. *Chem. Commun.* **47**, 7559–7566 (2011).
- Tobias, N. J., Shi, Y. M. & Bode, H. B. Refining the natural product repertoire in entomopathogenic bacteria. *Trends Microbiol.* **26**, 833–840 (2018).

- Tambong, J. T. Phylogeny of bacteria isolated from *Rhabditis* sp. (Nematoda) and identification of novel entomopathogenic *Serratia marcescens* strains. *Curr. Microbiol.* **66**, 138–144 (2013).
- Yokoyama, K. & Lilla, E. A. C–C bond forming radical SAM enzymes involved in the construction of carbon skeletons of cofactors and natural products. *Nat. Prod. Rep.* **35**, 660–694 (2018).
- Schramma, K. R., Bushin, L. B. & Seyedsayamdost, M. R. Structure and biosynthesis of a macrocyclic peptide containing an unprecedented lysine-to-tryptophan crosslink. *Nat. Chem.* **7**, 431–437 (2015).
- Embley, T. M. & Stackebrandt, E. The molecular phylogeny and systematics of the actinomycetes. *Annu. Rev. Microbiol.* **48**, 257–289 (1994).
- Lloyd-Price, J., Abu-Ali, G. & Huttenhower, C. The healthy human microbiome. *Genome Med.* **8**, 51 (2016).
- Bokulich, N. A. et al. Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci. Transl. Med.* **8**, 343ra82 (2016).
- O'Shea, R. & Moser, H. E. Physicochemical properties of antibacterial compounds: implications for drug discovery. *J. Med. Chem.* **51**, 2871–2878 (2008).
- Ling, L. L. et al. A new antibiotic kills pathogens without detectable resistance. *Nature* **517**, 455–459 (2015).
- Kononova, A., Kahne, D. E. & Silhavy, T. J. Outer membrane biogenesis. *Annu. Rev. Microbiol.* **71**, 539–556 (2017).
- Ghequire, M. G. K., Swings, T., Michiels, J., Buchanan, S. K. & De Mot, R. Hitting with a BAM: selective killing by lectin-like bacteriocins. *mBio* **9**, e02138-17 (2018).
- Storek, K. M. et al. Monoclonal antibody targeting the β -barrel assembly machine of *Escherichia coli* is bactericidal. *Proc. Natl Acad. Sci. USA* **115**, 3692–3697 (2018).
- Hart, E. M. et al. A small-molecule inhibitor of BamA impervious to efflux and the outer membrane permeability barrier. *Proc. Natl Acad. Sci. USA* **116**, 21748–21757 (2019).
- Hartmann, J.-B., Zahn, M., Burmann, I. M., Bibow, S. & Hiller, S. Sequence-specific solution NMR assignments of the β -barrel insertase BamA to monitor its conformational ensemble at the atomic level. *J. Am. Chem. Soc.* **140**, 11252–11260 (2018).
- Kaur, H. et al. Identification of conformation-selective nanobodies against the membrane protein insertase BamA by an integrated structural biology approach. *J. Biomol. NMR* **73**, 375–384 (2019).
- Ramos-Castañeda, J. A. et al. Mortality due to KPC carbapenemase-producing *Klebsiella pneumoniae* infections: systematic review and meta-analysis: mortality due to KPC *Klebsiella pneumoniae* infections. *J. Infect.* **76**, 438–448 (2018).
- Xu, L., Sun, X. & Ma, X. Systematic review and meta-analysis of mortality of patients infected with carbapenem-resistant *Klebsiella pneumoniae*. *Ann. Clin. Microbiol. Antimicrob.* **16**, 18 (2017).
- Sun, J., Zhang, H., Liu, Y. H. & Feng, Y. Towards understanding MCR-like colistin resistance. *Trends Microbiol.* **26**, 794–808 (2018).
- Levasseur, P. et al. Efficacy of a ceftazidime–avibactam combination in a murine model of septicemia caused by Enterobacteriaceae species producing ampc or extended-spectrum β -lactamases. *Antimicrob. Agents Chemother.* **58**, 6490–6495 (2014).
- Wunderink, R. G. et al. Effect and safety of meropenem–vaborbactam versus best-available therapy in patients with carbapenem-resistant Enterobacteriaceae infections: the TANGO II randomized clinical trial. *Infect. Dis. Ther.* **7**, 439–455 (2018).
- King, A. M. et al. Aspergillomarasmine A overcomes metallo- β -lactamase antibiotic resistance. *Nature* **510**, 503–506 (2014).
- Liu, J., Smith, P. A., Steed, D. B. & Romesberg, F. Efforts toward broadening the spectrum of arylomycin antibiotic activity. *Bioorg. Med. Chem. Lett.* **23**, 5654–5659 (2013).
- Smith, P. A. et al. Optimized arylomycins are a new class of Gram-negative antibiotics. *Nature* **561**, 189–194 (2018).
- Richter, M. F. et al. Predictive compound accumulation rules yield a broad-spectrum antibiotic. *Nature* **545**, 299–304 (2017).
- Crits-Christoph, A., Diamond, S., Butterfield, C. N., Thomas, B. C. & Banfield, J. F. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature* **558**, 440–444 (2018).
- Tobias, N. J. et al. Natural product diversity associated with the nematode symbionts *Photorhabdus* and *Xenorhabdus*. *Nat. Microbiol.* **2**, 1676–1685 (2017).
- Poinar, G. Jr. Origins and phylogenetic relationships of the entomophilic rhabditids, *Heterorhabditis* and *Steinernema*. *Fundam. Appl. Nematol.* **16**, 333–338 (1993).
- Bakelar, J., Buchanan, S. K. & Noinaj, N. The structure of the β -barrel assembly machinery complex. *Science* **351**, 180–186 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

Screening conditions

Photorhabdus and *Xenorhabdus* strains used in this study were purchased from Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ) or provided by H. Goodrich-Blair. Strains were inoculated in 10 ml Luria–Bertani (LB) broth in 50-ml Falcon tubes and incubated overnight, then diluted 1:100 in new Falcon tubes with 10 ml LB broth, nutrient broth or tryptic soy broth and incubated for 8 days, at 28 °C with shaking at 200 rpm. Culture aliquots (1 ml) were centrifuged at 12,000g for 10 min, and supernatants (750 µl) were collected and dried by centrifugal evaporation. Dried samples were resuspended in 50 µl Milli-Q water or 50% dimethyl sulfoxide to generate 15× concentrated extracts, then 3 µl was spotted onto *E. coli* overlays. Overlays were prepared from an exponential culture of *E. coli* (grown for 2–5 h after dilution of 1:100 from an overnight culture in cation-adjusted Mueller Hinton II broth (MHIIb) and incubated at 37 °C with shaking at 220 rpm), diluted to an optical density at 600 nm (OD_{600}) of 0.03 in MHIIb. These cultures were used to cover cation-adjusted Mueller Hinton II agar (MHIIa) plates; the excess culture was removed and overlays were left to dry in a biosafety cabinet. Overlays spotted with culture extracts were incubated at 37 °C overnight and the activity was evaluated by zones of inhibition.

Strain fermentation and purification of darobactin

P. khanii strains were inoculated in a 500-ml Erlenmeyer flask with 200 ml LB broth and incubated at 28 °C with aeration at 200 rpm overnight, then diluted 1:100 into a 2-l Erlenmeyer flask with 1 l tryptic soy broth and incubated for 10–14 days. Cells were removed by centrifugation at 8,000g for 10 min, and the culture supernatant was incubated overnight with XAD16N resin (20–60 mesh, Sigma-Aldrich), under agitation, to bind darobactin. Darobactin was eluted from the XAD16N resin using 1 l of 50% methanol with 0.1% formic acid. The eluate was concentrated using a rotary evaporator, and subjected to cation-exchange (SP Sepharose XL, GE Healthcare) chromatography. The concentrated eluate was loaded on to the activated cation-exchange resin and the resin washed with 0.1% (v/v) formic acid in ddH₂O. The compound was eluted by step gradients of 50 mM ammonium acetate pH 5, pH 7 and pH 8. The bioactive eluates were combined and freeze-dried, then resuspended in 0.1% (v/v) formic acid in Milli-Q water. The solution was subjected to reverse-phase high-performance liquid chromatography (RP-HPLC) on a C18 column (Agilent, C18, 5 µm; 250 mm × 10 mm, Restek). HPLC conditions were as follows: solvent A, Milli-Q water and 0.1% (v/v) formic acid; solvent B, acetonitrile and 0.1% (v/v) formic acid. The initial concentration of 2% solvent B was maintained for 2 min, followed by a linear gradient to 26% over 12 min with a flow rate of 5 ml min⁻¹; UV detection by diode-array detector from 210 to 400 nm. Darobactin was eluted at 12.5 min, with a purity of 97% by UV.

Structure elucidation

Mass spectrometric analysis. The exact mass of darobactin was determined using a Q Exactive Hybrid Quadrupole-Orbitrap Mass Spectrometer (Thermo Scientific) equipped with a heated electrospray ionization source operated in positive ionization mode. Darobactin was prepared in Milli-Q water and 0.1% formic acid and introduced into the mass spectrometer by direct infusion at a constant flow rate of 5 µl min⁻¹. The ion source conditions were set as follows: ion spray voltage, 1.50 kV; capillary temperature, 125 °C; spray current, 50 µA; sheath gas, 0; and aux gas, 2. The tandem mass spectrometry (MS/MS) spectrum for darobactin was acquired in higher-energy collisional dissociation mode and a collision energy of 55 eV was applied for the fragmentation. The mass analyser was calibrated according to the manufacturer's directions. Data acquisition and processing were performed using Xcalibur software (Thermo Fisher Scientific).

NMR studies. All NMR data were recorded on a Bruker AVANCE II 700-MHz NMR spectrometer with 5 mm TXI probehead and a Bruker AVANCE NEO 600-MHz NMR spectrometer equipped with a 5 mm TCI cryoprobe. Complete assignments were obtained using two-dimensional experiments, including COSY (cosydfesgpph), TOCSY (dipsi2esfbgpph), ¹H-¹⁵N HSQC (hsqcetfp3gpsi), ¹H-¹³C HSQC (hsqcetgpsi2.3), ¹H-¹³C HMBC (hmbcgp1pndprqf) and ROESY (roesyegpph). All NMR experiments were performed with 5 mg of darobactin solubilized in 500 µl of aqueous solvent containing 94% (v/v) Milli-Q water, 4% (v/v) deuterium oxide and 2% (v/v) deuterated formic acid. Additional ¹D-¹H and ²D HMBC and ROESY NMR experiments were performed with 5 mg of darobactin solubilized in 500 µl of 2:1 (v/v) mixture of Milli-Q water and deuterated acetonitrile, including 2% (v/v) deuterated formic acid.

Modelling of isomers. Modelling of the four possible darobactin isomers was performed in Schrodinger 2018-2. The four isomers first underwent conformational search in the MacroModel module (Schrödinger) with MMFF forcefield. Mixed torsional/low-mode sampling method was used with a maximum of 10,000 steps. The lowest energy conformer for each isomer was then subjected to geometry optimization using Jaguar (Schrödinger) at B3LYP/6-31G (d, p) level with fine-grid density and the ultrafine accuracy level of SCF. All simulations were performed for gas phase.

Identification of the BGC

The genome of *P. khanii* HGB1456 was sequenced by both Pacbio technology and Illumina Miseq, and assembled using SPAdes 3.11³⁶. The resulting data were initially analysed using antibiotic and secondary metabolite analysis shell (antiSMASH³⁷). Each predicted BGC was then analysed manually, taking into account the number and identity of predicted amino acids. As this initial approach did not yield any putative darobactin BGCs, a direct screening for the core peptide sequence WNWSKSF was done on all *Photorhabdus* genomes available in public databases using the Basic Local Alignment Search Tool (BLAST). In *P. khanii*, the seven amino acid sequence of darobactin was located close to the C terminus of an open-reading frame that encodes 58 amino acids, upstream of an ABC transporter and a radical SAM enzyme, suggesting a RiPP operon. This putative BGC was identified in the other darobactin producers *P. luminescens* DSM3368 and *P. khanii* DSM3369. The boundaries of the cluster were determined by comparison with the *P. bodei* genome, which did not contain the operon. Furthermore, the GC content of the *dar* cluster was clearly lower than the rest of the average GC content in the genome (32% versus 45%).

To identify other bacterial species that potentially produced darobactin-like compounds, homologous enzymes were searched using the radical SAM protein sequence (DarE) as input in BLAST. The genomic context of each hit was analysed manually to confirm the presence of a DarA-like propeptide in the vicinity of the radical SAM protein. In addition, a search using the propeptide DarA as input was done, delivering the same hits.

Generation of a darobactin deletion mutant and heterologous expression

To delete the *dar* BGC (*darABCDE*) from the genome of the producer strain *P. khanii* DSM3369, a plasmid was constructed by assembly of five fragments, which enables markerless genome modification. Chromosomal DNA was isolated using the innuprepBacteria DNA Kit (AnalytikJena). Fragments up- and downstream of the BGC were amplified (size of around 1 kb) using the primer pairs 5'-TTTGACGTGGAGTCCACGTGTTATGGACGTGGCAAACGCGGTTCTTGAC-3', 5'-TTGAAATATCAGGATAGCATTCGCTCGCTACCCCGGTCACATAGTTCG-3'; and 5'-ATGCTATCCTGATATTTCAAATGCAAGTAAATGTTTCATCATAATAACC-3' and 5'-TTCTTGACGAGTCTTCTGAGATGGGTGATATCCACTGATATAAATCTC-3'. Then, the R6K origin of replication (ori), the origin of transfer

Article

(oriT) and the levan sucrase gene *sacB* from *Bacillus subtilis* were amplified in one piece from the vector pNPTS138³⁸ using the primers 5'-TCGAGCTCTAAGGAGGTTATAAAAAATGAACATCAAAAAGTTTGCAAACAAGCA-3' and 5'-ACGTGGACTCCAACGTCAAA-3'. Next, the arabinose-inducible expression system of pKD46³⁹ with the adjacent β -lactamase (*bla*) promoter was amplified using the primers 5'-ACTCTTCCTTTTCAATATTATGAAGCAT-3' and 5'-TGCATTTTATAACCTCCTTAGAGCTCGAATTCC-3'. Finally, the *aph* gene from pCAP03⁴⁰, which confers resistance to kanamycin, was amplified using the primers 5'-TCAGAAGAACTCGTCAAGAAGGCGA-3' and 5'-TCAATAATATTGAAAAAGGAAGAGTATGATTGAACAAGATGGATTGACG-3'. All fragments were amplified by Q5 DNA polymerase (New England Biolabs), the gel was purified with 1% or 2% TAE agarose gels and the DNA was retrieved with the Large Fragment DNA Recovery Kit (Zymo Research). Subsequently all fragments were fused by isothermal assembly, generating the plasmid pNB02.

After assembly, *E. coli* WM3064 cells were transformed with pNB02 by electroporation and correct assembly was corroborated by PCR and restriction analysis following standard procedures. Conjugation between *E. coli* WM3064 and *P. khanii* DSM3369 was performed by growing both strains to an OD₆₀₀ of around 0.6. After washing twice with LB medium, cells were mixed in a 1:3 ratio of *E. coli* and *P. khanii*, plated onto LB agar supplemented with diaminopimelic acid (0.3 mM) and incubated at 37 °C for 3 h, followed by overnight incubation at 30 °C. The bacterial lawn was resuspended in LB medium and plated on LB agar with kanamycin (50 µg ml⁻¹) as a serial dilution. Kanamycin-resistant single crossover transconjugants were grown in LB medium to an OD₆₀₀ of approximately 0.6. Then, expression of *SacB* was induced by adding arabinose (0.2% w/v), followed by 2 h incubation. Subsequently, the culture was plated onto LB agar supplemented with 0.2% (w/v) arabinose and 10% sucrose and incubated at 30 °C for 48 h. Single colonies were picked on LB_{Kan} and LB_{Ara/Suc} agar. Sensitivity to kanamycin indicated plasmid loss and therefore a successful double crossover event. Clones were picked and analysed for BGC loss by PCR using the primers 5'-ATCTCCATCAAAGCGCTACC-3' and 5'-CCGCGTGCACCTCGAAATC-3'. The knockout strain is called *P. khanii* DSM3369 Δ *darABCDE*.

For heterologous expression of the darobactin ABGC in *E. coli* and to complement *P. khanii* DSM3369 Δ *darABCDE*, the expression plasmid pNB03 was used. To avoid issues with the regulation system between the propeptide and the modifying enzymes, all intergenic regions were removed and the genes *darA*–*darE* were expressed streamlined under the control of the arabinose-inducible *araB* promoter.

pNB03 was created by amplification of the p15A ori from pACYC177 (primers 5'-GGTCGACGGATCCCCGGAATAGCGGAATGGCTTACGAAC-3' and 5'-CTCTAAGGAGGTTATAAAAAGCGGCCGCATCCCTTAACGTGAGTTTTC-3'); the arabinose expression system and kanamycin-resistance gene of pNB02 (primers 5'-AAGCAGCTCCAGCCTACATCAGAAGAACTCGTCAAGAAGGCGA-3' and 5'-TTTTTAACCTCCTTAGAGCTCGAATTCC-3'), oriT and the *aac(3)* gene, which confers resistance to apramycin from pIJ773⁴¹ (primers 5'-ATTCCGGGATCCGTCGACC-3' and 5'-TGTAGGCTGGAGCTGCTT-3'). Subsequently, all fragments were gel purified and assembled as described previously. *E. coli* TOP10 cells were transformed with the vector and correct assembly was corroborated. To introduce the *dar* BGC into *P. khanii* DSM3369 Δ *darABCDE*, pNB03 was first linearized using the primers 5'-TCCCTTAACGTGAGTTTTCG-3' and 5'-TTTTATAACCTCCTTAGAGCTCGAA-3', *darA* was then amplified using 5'-GCTCTAAGGAGGTTATAAAAATGCATAATACCTTAAATGAAACCGTTAAA-3' and 5'-AATAGCATTCATTTATGGCTCTCCTTTTAAATTCCTGGAAGCTTT-3', and *darB*–*darE* was amplified using 5'-AAAGCTTCCAGGAAATTTAAAAGGAGAGCCATAATGAATGCTATT-3' and 5'-CGAAAACCTACGTTAAGGGATTACGCCGCGATGGTTGTTTATT-3'. All fragments were gel purified and assembled as described above. The resulting vector pNB03-*darABCDE* was transferred to *E. coli* TOP10 cells and correct assembly was corroborated.

The empty pNB03 and pNB03-*darABCDE* vectors were transferred to *P. khanii* DSM3369 Δ *darABCDE* by triparental conjugation. In brief, conjugation between *P. khanii* DSM3369 Δ *darABCDE*, *E. coli* TOP10 carrying the expression plasmid and *E. coli* ET pUB307, which carried the pUB307 conjugation helper plasmid, was carried out as described above (cell ratio 3:1:1). As *P. khanii* DSM3369 is naturally resistant to carbenicillin and the kanamycin resistance of pUB307 lacks the *bla* promoter, final selection took place on LB agar supplemented with kanamycin and carbenicillin. Kanamycin-resistant transconjugants were grown in LB_{Kan}, the plasmid was isolated and the identity verified by PCR. For heterologous expression, the vector pNB03-*darABCDE* was transferred to *E. coli* BW25113 (arabinose non-utilizer) by electroporation.

Subsequently, wild-type *P. khanii* DSM3369, *P. khanii* DSM3369 Δ *darABCDE* and pNB03, *P. khanii* DSM3369 Δ *darABCDE* and pNB03-*darABCDE*, and *E. coli* and pNB03-*darABCDE* were grown in LB or LB_{Kan} supplemented with 0.2% (w/v) arabinose for 5–7 days and analysed by liquid chromatography coupled to mass spectrometry (LC–MS).

Then, the centrifuged culture supernatant was desalted on self-packed C18 columns by washing with 5% acetonitrile, and subsequent elution with 80% acetonitrile in Milli-Q water and 0.1% formic acid. A Dionex UltiMate 3000 HPLC system was coupled to a high-resolution electrospray ionization quadrupole time-of-flight mass spectrometer (QqTOF-ESI-HRMS) from Bruker Daltonics Instruments. Dionex Acclaim 120 C18 (5 µm 4.6 mm × 100 mm) was used for the separation with solvent A (Milli-Q water) and solvent B (100% methanol). The initial concentration of 10% solvent B was maintained for 10 min, followed by a linear gradient to 100% over 30 min. MS parameters were as follows: nebulizer gas, 1.6 bar; gas temperature, 200 °C; gas flow, 8 l min⁻¹; capillary voltage, 4,500 V; endplate offset, 500 V; positive ion mode.

MIC

The MIC was determined by microbroth dilution. Under aerobic conditions, overnight cultures of *E. coli* strains, *P. aeruginosa* strains, *A. baumannii* ATCC 17978, *K. pneumoniae* strains and *S. aureus* HG003, were diluted 1:100 in MHIIB and incubated at 37 °C with aeration at 220 rpm. Exponential cultures (OD₆₀₀ of 0.1–0.9) were diluted to an OD₆₀₀ of 0.001 (approximately 5 × 10⁵ c.f.u. ml⁻¹) in MHIIB and 98 µl aliquots were transferred into round-bottom 96-well plates containing 2 µl of darobactin solutions diluted serially twofold. After overnight incubation at 37 °C, the darobactin MIC was determined as the minimum concentration at which no growth of strains could be detected by eye. For susceptibility testing of *Mycobacterium tuberculosis*, cells were cultured in BD Difco 7H9 base medium supplemented with 10% OADC enrichment (oleic acid, albumin, dextrose and catalase), 0.5% glycerol, 0.2% casamino acids, 0.05% tyloxapol, 80 µg ml⁻¹ lysine and 24 µg ml⁻¹ pantothenate. An exponentially growing culture of strain m²6020 (Δ *lysA* Δ *panCD*) was diluted to an OD₆₀₀ of 0.003 (approximately 5 × 10⁵ cells ml⁻¹) and seeded into 96-well plates containing darobactin dilutions. The plates were incubated for 5 days, then resazurin was added to each well to a final concentration of 2.5 µg ml⁻¹. The plates were incubated for an additional 2 days, at which point the MIC was determined by eye. The MIC against intestinal pathobionts and symbionts (*Shigella sonnei*, *Salmonella enterica* Typhimurium LT2, *Moraxella catarrhalis*, *Enterobacter cloacae*, *Bifidobacterium longum*, *Bacteroides fragilis* and *Lactobacillus reuteri* (ATCC 25931, 19585, 25238, 13047, BAA-999, 25285 and 23272, respectively); KLE collection bacteria were isolated from stool under anaerobic conditions and identified by 16S sequencing) was determined under anaerobic conditions (Coy Vinyl Anaerobic chamber, 37 °C, 5% H₂, 10% CO₂, 85% N₂). Overnight cultures grown in brain–heart infusion (BHI) broth, supplemented with 0.5% yeast extract, 0.1% L-cysteine hydrochloride and 15 µg ml⁻¹ haemin (BHI-Ych), were diluted 1:100 in BHI-Ych. The 96-well assay plates were prepared by twofold dilution of darobactin, and included a positive growth control. After 24 h incubation, the MIC was determined. All MIC assays were performed at least in triplicate. The MIC against clinical isolates

(Supplementary Table 1) of *E. coli*, *K. pneumoniae* and *P. aeruginosa* was evaluated by JMI laboratories.

Cytotoxicity

A microplate Alamar blue assay (MABA/resazurin) was used to determine the cytotoxicity of darobactin. Exponentially growing FaDu pharynx squamous cell carcinoma (ATCC HTB-43), HepG2 liver hepatocellular carcinoma (ATCC HB-8065) and red-fluorescent-protein (RFP)-tagged human embryonic kidney 293 (HEK293-RFP; GenTarget SC007) cells, all cultured in Eagle's minimum essential medium supplemented with 10% fetal bovine serum were seeded into a 96-well, flat-bottom, tissue-culture-treated plate (Corning) and incubated at 37 °C with 5% CO₂. After 24 h, the medium was aspirated and replaced with fresh medium containing test compounds (2 µl of a twofold serial dilution in water to 98 µl of medium). After 72 h of incubation at 37 °C with 5% CO₂, resazurin (Acros Organics) was added to each well to a final concentration of 0.15 mM. After 3 h, the absorbances at 544 nm and 590 nm were measured using a BioTek Synergy H1 microplate reader. Experiments were performed in biological triplicate.

Time-dependent killing

An overnight culture of *E. coli* MG1655 was diluted 1:10,000 in MHIIB and incubated at 37 °C for 2 h with aeration at 220 rpm. *E. coli* was treated with 16× MIC antibiotic (64 µg ml⁻¹ darobactin and 64 µg ml⁻¹ ampicillin) and the time at which each antibiotic was added was defined as 0 h. At each time point, 100-µl aliquots were collected and centrifuged at 12,000g for 5 min, pellets washed with 100 µl PBS and resuspended in 100 µl PBS and tenfold serially diluted suspensions were plated onto MHIIA. After overnight cultivation at 37 °C, colonies were counted and c.f.u. ml⁻¹ was calculated. Experiments were performed in biological triplicate.

Resistance studies

E. coli MG1655 cells from an exponential culture were washed in PBS, and subsequently inoculated onto 30 MHIIA plates containing 4× MIC darobactin, at a density of 5 × 10⁷ c.f.u. per plate. After 2 days of cultivation at 37 °C, plates were examined for colonies, the number of colonies was counted and the colonies were restreaked to test for resistance stability. Subsequently, the colonies were tested by 16S sequencing to ensure that the colonies were *E. coli*. To evolve resistance to darobactin in liquid culture, an overnight culture of *E. coli* MG1655 was diluted 1:100 in 1 ml MHIIB containing 0.5×, 1×, 2× and 4× MIC darobactin and incubated at 37 °C for 24 h with aeration at 220 rpm. The darobactin concentration that inhibited growth of *E. coli* below an OD₆₀₀ of 2.0 was defined as the MIC, and the culture at 0.5× MIC (OD₆₀₀ > 2) was used to re-inoculate tubes with 0.5×, 1×, 2× and 4× the new MIC at 1:100. This was repeated until cultures were able to grow in 256 µg ml⁻¹ darobactin, and cultures were then maintained in 256 µg ml⁻¹ darobactin until the end of the experiment. Experiments were performed with three independent cultures. For the mutation analysis, more than 3 million paired-end Illumina reads were sequenced for each resistant mutant and mapped to the *E. coli* MG1655 genome (GenBank accession U00096.3) using Geneious v.11.0.4. Single-nucleotide polymorphisms were called using the default parameters, and the generated variant call format (VCF) files were manually filtered to remove calls with a quality score of less than 1,000.

Scanning electron microscopy

E. coli MG1655 samples were prepared as for the time-dependent killing experiments. After washing the cells with PBS, 10 µl cell suspensions were spotted onto Aclar film coated with 0.1% poly-L-lysine. *E. coli* cells were fixed with 2.5% glutaraldehyde in 0.1 M sodium cacodylate containing 0.15% Alcian blue and 0.15% safranin O for 24 h at 4 °C. The samples were washed in 0.1 M sodium cacodylate for 5–10 min, infiltrated with 1% osmium tetroxide for 30 min, washed three times in 0.1 M sodium

cacodylate, and then dehydrated by a graded series of ethanol concentrations (30%, 50%, 70%, 85%, 95% and 100%) for 5–10 min for each concentration. The dehydration step with 100% ethanol was repeated three times. Critical point drying was performed using SAMDRI-PVT-3D (Tousimis) from liquid CO₂. The samples were mounted onto an aluminium sample mount using double-sided conductive-carbon adhesive tape and coated with 5 nm platinum by sputter coating (Cressington 208HR). The samples were imaged with Hitachi S-4800 (Hitachi) at 3.0 kV.

Fluorescence microscopy

E. coli MG1655 was cultured in MHIIB until stationary phase, inoculated into fresh MHIIB at 1:10,000 and grown for 2 h at 37 °C. Cells were concentrated 50-fold in MHIIB, placed on top of a MHIIB–darobactin (64 µg ml⁻¹) 1.5% low-melting agarose pad containing FM4-64 (10 µg ml⁻¹) and Sytox Green (0.5 µM) dyes from Molecular Probes and observed using a ZEISS LSM 710 confocal microscope using a 63× oil-immersion objective lens. The two signals from FM4-64 and Sytox Green were collected after excitation at 488 nm, alongside a differential interference contrast image. The differential interference contrast, FM4-64 and Sytox Green signals were acquired every 30 min at a temperature of 37 °C maintained through a thermostatic chamber. Images were acquired by Zen Software at a resolution of 1,024 × 1,024 and lane average of 8, and processed with Fiji software⁴². The images shown in Extended Data Fig. 6 were processed using the enhance contrast process, and the HyperStackReg plugin was used to correct for the x–y drift in Supplementary Video 1.

Lipopolysaccharide binding assay

The lipopolysaccharide (LPS) binding assay was performed based on the MIC assay. Aliquots (100 µl) of *E. coli* MG1655 cultures with an OD₆₀₀ of 0.001 and grown in MHIIB were transferred into a 96-well plate containing purified LPS from *E. coli* O55:B5 (0.5–100 µg ml⁻¹; Sigma, L4524) and darobactin or polymyxin B. The antibiotic concentrations that inhibited *E. coli* MG1655 growth were determined in the absence or presence of LPS.

Construction of *bamA* recombinant mutant in *E. coli* MG1655 and ATCC 25922

The linear DNA product comprising the mutated *bamA* gene (1300A>G, 1334A>C and 2113G>A) was amplified by PCR, using the primers *bamA*-recF (5'-ACTATCTGGATCGCGGTTATGC-3') and *bamA*-recR (5'-TTCACAGCAGTCTGGATACGAG-3'), and the genomic DNA from *E. coli* darobactin-resistant mutant (strain 3) template. Approximately 500 ng of column-purified mutated *bamA* product was used to transform electrocompetent cells of *E. coli* MG1655-pKD46 to perform λ Red recombination^{39,43}. The subsequent steps have been adapted from the 'Quick and Easy *E. coli* Gene Deletion Kit' (GeneBridges). In brief, 30 µl of an overnight culture of *E. coli* MG1655-pKD46 was used to inoculate a microtube containing 1.4 ml of LB medium complemented with ampicillin (100 µg ml⁻¹). After 2 h of shaking at 30 °C, 0.4% of L-arabinose was added, and the tube was transferred for shaking at 37 °C for 1 h. Cells were washed and concentrated with ice-cold 10% glycerol before electroporation. The recovery step was performed for 3 h at 37 °C with shaking. First selection was performed using resistance to darobactin (32 µg ml⁻¹). Several transformant clones were then restreaked with double selection for resistance to darobactin (32 µg ml⁻¹) and sensitivity to ampicillin (100 µg ml⁻¹, at 30 °C). The *bamA* locus was amplified and the presence of the mutations (1300A>G, 1334A>C and 2113G>A leading to T434A, Q445P and A705T, respectively) was confirmed by sequencing.

For virulence testing, to transfer mutations from strain 3 into *E. coli* ATCC 25922 leading to the triple *bamA* mutant, the same strategy was used. During the manipulation of *E. coli* ATCC 25922 to construct the *bamA* recombinant mutant, two spontaneous *bamA* mutants with single

single-nucleotide polymorphisms were isolated from darobactin-containing plates ($16 \mu\text{g ml}^{-1}$): 1285G>A, leading to G429R and 1286G>T, leading to G429V.

Transcriptome analysis

For the challenge experiments, 3 ml of *E. coli* BW25113 at an OD_{600} of 0.5, representing mid-log phase, was exposed to $4 \mu\text{g ml}^{-1}$ darobactin for 15 min, 30 min and 1 h in biological triplicate. After exposure, cells were immediately pelleted at 4°C by centrifugation for 2 min at 2,000 rpm in 1-ml aliquots. The supernatants were removed and samples were immediately frozen in liquid nitrogen at -80°C until they were processed for total RNA isolation. Total RNA was extracted by automation using the NucleoMag RNA extraction kit on the EpMotion Robotic liquid handler. For the resulting total RNA, RIN values were obtained to check for RNA quality using the 2200 TapeStation instrument from Agilent Genomics. rRNAs were subtracted from the total RNA to yield only mRNA for library construction using the NEB bacterial rRNA depletion kit at half reactions with a total RNA input maximum of 400 ng. The quality of the rRNA-depleted samples was checked using an Agilent Bioanalyzer with the Agilent Pico chip for RNA detection for less than 0.5% of rRNA remaining in each sample. Then, 2–5 ng of the rRNA-depleted samples was used as the input material to construct each cDNA library for RNA sequencing using the NEBNext Ultra Directional RNA Library prep kit from Illumina. The quality of the resulting libraries was checked using Agilent High Sensitivity DNA chips to ensure proper library size distribution and the absence of small adapters. Libraries were quantified and normalized by qPCR and then sequenced using the NextSeq 500 High Output Kit at 150 cycles producing approximately 9 million, 75-bp, paired-end reads for each library. These reads were mapped to *E. coli* strain BW25113 using *clc_assembler* v.4.4.2.133896 (CLC Genomics Workbench 11.0). Differential expression was computed using *edgeR::exactTest*⁴⁴ in R v.3.5.1 with unnormalized gene counts ($n = 4,626$ genes) for each treatment at time t versus the matched control, for $t \in \{15, 30, 60\}$. The gene count matrix was restricted to genes at minimum present in all replicates from one treatment condition resulting in $n = 4,514$ genes. Volcano plots were created using *plot_volcano* from *soothsayer* (<https://github.com/jolespin/soothsayer>) in Python v.3.6.6. Directed networks (DiNetwork) were constructed and plotted using NetworkX and Matplotlib Python packages, respectively. Heat maps were generated using *seaborn* and operon plots were created with Matplotlib.

Two-dimensional thermal proteome profiling

Thermal proteome profiling was performed as previously described^{45,46}. In brief, *E. coli* BW25113 cells were grown aerobically at 37°C with shaking until an OD_{578} of approximately 0.7 was reached. For living-cell experiments, darobactin was then added at five different concentrations and incubated for 10 min. For experiments in which protein synthesis was inhibited, cells were treated with 0.2 mg ml^{-1} chloramphenicol for 10 min before addition of darobactin. For lysate experiments, cells were disrupted with five freeze–thaw cycles before darobactin treatment. Aliquots of treated cells or lysates were then heated for 3 min to 10 different temperatures in a PCR machine (Agilent SureCycler 8800). After cell lysis, protein aggregates were removed and the remaining soluble proteins were digested according to a modified SP3 protocol^{47,48}, as previously described⁴⁹. Peptides were labelled with TMT10plex (ThermoFisher Scientific), fractionated onto six fractions under high pH conditions and analysed using LC–MS/MS, as previously described⁴⁵. Protein identification and quantification was performed using *IsobarQuant*⁵⁰ and *Mascot* 2.4 (Matrix Science) against the *E. coli* Uniprot FASTA (Proteome identifier, UP000000625). Data were analysed with the TPP package for R⁵⁰ followed by a false-discovery rate (FDR)-controlled method for functional analysis of dose–response curves⁴⁹. Data are available in Supplementary Table 2.

Cloning, expression and purification of BAM and BAM mutants for nanodiscs

To prepare the BAM mutants, the pJH114 plasmid (a gift from H. Bernstein)³⁵ was used as a template using an Agilent QuikChange Lightning Multi Site-Directed Mutagenesis Kit (Agilent); oligonucleotide sequences are available upon request. The plasmids encoding wild-type BAM (pJH114), with BamE carrying a C-terminal His-tag, and the corresponding BamA mutant genes 1–3 (M1, G429V and G807V; M2, F394V, E435K and G443D; M3, T434A, Q445P and A705T) were cloned under an isopropyl- β -D-thiogalactoside (IPTG) promoter and sequences were confirmed. The primers used for mutation are as below: BAM_mutation1_G429V (5'-TTCAACTTTGTTATTGGTTAC-3'), BAM_mutation1_G807V (5'-TTTAACATCGTTAAACCTGG-3'), BAM_mutation2_F394V (5'-CGTCTGGGCGTCTTTGAAAC-3'), BAM_mutation2_E435K (5'-TACGGTACTAAAAGTGGCGTG-3'), BAM_mutation2_G443D (5'-TTCCAGGCTGATGTGCAGCAG-3'), BAM_mutation3_T434A (5'-GGTTACGGTGCTGAAAGTGGC-3'), BAM_mutation3_Q445P (5'-GCTGGTGTGCCGAGGATAAC-3') and BAM_mutation3_A705T (5'-TCGGATGATACTGTAGGCGG-3'). Plasmids were transformed into BL21 (DE3) cells, plated onto LB–carbenicillin agar plates and incubated overnight at 37°C . After transforming the plasmids into *E. coli* BL21 (DE3), the plasmids were isolated and resequenced. The sequence of M2 and M3 was unchanged, but an additional mutation, T434A, appeared in M1, which we refer to as M1a. This additional mutation matches the T434A mutation in M3, and may have been selected for during cell growth, possibly stabilizing the protein. A 50-ml overnight culture was prepared from a single colony in $2\times$ YT medium supplemented with $100 \mu\text{g ml}^{-1}$ of ampicillin. The cells were then centrifuged and resuspended in 5 ml of fresh $2\times$ YT medium and then 1 ml was added to five 2-l baffled flasks containing 1 l of $2\times$ YT medium supplemented with $50 \mu\text{g ml}^{-1}$ of ampicillin. These cultures were grown at 37°C until an OD_{600} between 0.8 and 1.0 was reached, after which the cultures were induced with 0.5 mM IPTG and the cells were collected after 3 h. Purification was performed as previously described³⁵. In brief, cells were resuspended in lysis buffer ($1\times$ PBS, $10 \mu\text{g ml}^{-1}$ DNase I, $200 \mu\text{M}$ PMSF, $2 \mu\text{M}$ leupeptin and 1.5 nM pepstatin A) and lysed by three passages through an Emulsiflex C-3 homogenizer (Avestin) at 18,000 psi. The lysate was then centrifuged at $6,000g$ for 20 min and the supernatant was centrifuged at $200,000g$ for 90 min at 4°C . The membrane pellets were resuspended into solubilization buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.5% *n*-dodecyl- β -D-maltoside (DDM) and 37 mM imidazole) using a Dounce homogenizer, and subsequently stirred at 4°C for 4 h. Solubilized membranes were then centrifuged at $200,000g$ for 40 min at 4°C to collect the supernatant, which was used for purification using a 5-ml Ni-NTA column on an ÄKTA system (GE Healthcare) using buffer A (25 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.05% DDM and 37 mM imidazole) and buffer B (25 mM Tris-HCl, pH 7.5, 150 mM NaCl, 0.05% DDM and 1 M imidazole). Fractions containing the protein were pooled, concentrated to 5 ml for size-exclusion chromatography using a 16/60 Sephacryl S-300 HR column at a flow rate of 1.0 ml min^{-1} in 25 mM Tris-HCl, pH 7.5, 150 mM NaCl and 0.6% C_8E_4 . The peak fractions were pooled and concentrated as necessary.

Reconstitution of the BAM into nanodiscs

The membrane scaffold protein MSP1E3D1 was expressed and purified from *E. coli* as previously described^{51,52}. BAM was purified by size-exclusion chromatography in 25 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1.0% *n*-octyl- β -D-glucopyranoside and concentrated to $100 \mu\text{M}$. Nanodisc reconstitution was performed in a final volume of 300 μl by adding 20 μM of purified BAM, 100 μM of MSP1E3D1 and 2 mM of *E. coli* polar lipids (Avanti Polar Lipids) to a buffer containing 25 mM Tris-HCl, pH 7.5 and 150 mM NaCl. Bio-beads SM2 (Biorad) were added to the mixture and incubated at 4°C overnight. The Bio-beads were spun down and the supernatant was incubated with 300 μl of HisPur Ni-NTA Resin

(ThermoFisher Scientific) for 30 min at 4 °C. The BAM-inserted nanodiscs were then eluted from the Ni-NTA resin with 25 mM Tris-HCl, pH 7.5, 150 mM NaCl and 400 mM imidazole. The elution was then loaded onto a Superdex 200 Increase 10/300 GL column (GE Healthcare) in 25 mM Tris-HCl, pH 7.5 and 150 mM NaCl. The peak fractions were then pooled and concentrated to 40 µM.

BAM folding assay

OmpT and SurA (periplasmic chaperone) were expressed and purified from *E. coli* as previously reported^{53,54}. Solution 1 contained 0.4 µM BAM-nanodiscs, 0.6 µM of the fluorogenic peptide (Abz-Ala-Arg-Arg-Ala-Tyr(NO₂)-NH₂), and 0.1 mg ml⁻¹ LPS in 25 µl of 20 mM Tris-HCl, pH 6.5. Empty nanodiscs were used as a negative control. Solution 2 contained 20 µM urea-denatured OmpT with 140 µM SurA in 25 µl of 20 mM Tris-HCl, pH 6.5. To initiate the BAM folding reaction, solution 2 was incubated at room temperature for 10 min and then mixed with solution 1. Darobactin was added to solution 1 and incubated for 10 min before being mixed with solution 2. The fluorescence signal was monitored at 430 nm (excitation at 325 nm) using a SpectraMax M2e fluorescent plate reader (Molecular Devices) for 60 min with readings every 8 s. Data were then analysed and plotted using the online IC₅₀ Calculator tool (AAT Bioquest) and GraphPad Prism v8.2.

Isothermal titration calorimetry

All isothermal titration calorimetry (ITC) experiments were carried out at 25 °C with the NanoITC microcalorimeter (TA Instruments) in duplicate. BAM (300 µl) at a concentration of 20 µM in 1× PBS supplemented with 0.05% DDM was placed in the sample cell, and the ligand (darobactin or the linear peptide) with a concentration of 200 µM in the syringe (50 µl) was injected in 20 successive injections with a spacing of 300 s and a stirring rate of 300 rpm. Control experiments in the absence of BAM were performed under identical conditions to determine the heat signal attributable to only the injection of the ligand to the buffer. The resulting data were analysed and fit to the independent binding model using the NanoAnalyze software package (TA Instruments).

Sample preparation of BamA-β in lauryldimethylamine-*N*-oxide micelles for NMR

The protein construct comprising the β-barrel of *E. coli* BamA (residues 426–810, C690S, C700S; termed BamA-β) was established previously and sample production followed published protocols²¹. In brief, protein expression was carried out in *E. coli* BL21 (DE3) Lemo cells in M9 medium containing ¹⁵NH₄Cl and D₂O. Once the OD₆₀₀ reached 0.8, expression into inclusion bodies was induced by 1 mM IPTG at 37 °C for 12 h. The collected cells were resuspended in buffer A (20 mM Tris pH 8.0 and 300 mM NaCl) and lysed by sonication. Inclusion bodies were collected by centrifugation at 30,000g for 1 h and solubilized into 20 mM Tris pH 8.0 and 6 M guanidinium hydrochloride for 2 h. The solubilized sample was loaded onto Ni-NTA beads preequilibrated with buffer A supplemented with 6 M guanidinium hydrochloride. The protein was eluted with buffer A, containing 6 M guanidinium hydrochloride and 200 mM imidazole. Refolding was carried out in 20 mM Tris, 150 mM NaCl, pH 8.0 and 0.5% w/v lauryldimethylamine-*N*-oxide (LDAO) at 4 °C. The refolded sample was dialysed in 20 mM Tris, pH 8.0 overnight. Afterwards, folded BamA-β was purified by ion exchange in 20 mM Tris pH 8.0, 0.1% LDAO and the protein was eluted with a linear gradient of 0.5 M NaCl. Finally, BamA-β was loaded onto a size-exclusion chromatography column (HiLoad 16/600 Superdex 200 pg, GE Healthcare) in 20 mM phosphate buffer pH 7.5, 150 mM NaCl and 0.1% LDAO yielding a monomeric sample.

Solution NMR spectroscopy

A sample was concentrated to an initial protein concentration of 250 µM. Darobactin was added stepwise from a stock solution to 0.5:1-, 1:1- and 2:1-fold stoichiometry darobactin:BamA-β. At each

titration step, a two-dimensional [¹⁵N, ¹H]-TROSY experiment with 64 transients was recorded on a 700-MHz Bruker spectrometer equipped with a cryogenic probe at 37 °C. Then, 1,024 and 128 complex points were acquired in the direct and indirect dimension, respectively, and zero-filled to 2,048 and 256 points during processing. As a control experiment, a linear scrambled peptide WNKWSFS was synthesized, and added at 230 µM to BamA-β. The NMR spectra of 0:1-, 1:1- and 2:1-fold stoichiometry with darobactin:BamA-β, and 0:1 and 1:1 with the peptide WNKWSFS are provided as raw data (Supplementary Data 1). From these, spectra shown in Fig. 3c and Extended Data Fig. S1, j have been produced. The data format is readable using the standard NMR software TOPSPIN 3.6.2. An upper limit estimate for the dissociation constant *K_d* was obtained from a quantification of the relative amounts of ligand-free and ligand-bound BamA from NMR signal intensities under consideration of the spectral noise (Fig. 3c).

Animal studies

All animal studies were performed at Northeastern University, approved by Northeastern IACUC, and were performed according to institutional animal care and use policies. Experiments were not randomized nor blinded, as it was not deemed necessary. Female CD-1 mice (20–25 g, experimentally naive, 6 weeks old) from Charles River were used for all studies.

Virulence model. *E. coli* ATCC 25922, both wild-type and with *bamA* mutations leading to darobactin resistance (see ‘Construction of *bamA* recombinant mutant in *E. coli* MG1655 and ATCC 25922’), were tested in an acute infection model. An overnight culture (OD₆₀₀ of 2.0) of *E. coli* was diluted 1:10 in MHIB. Mice were infected with 0.1 ml of bacterial suspension, 2 × 10⁷ c.f.u. for all strains (determined by plate counts), and monitored for survival. At 24 h after infection, mice were euthanized by CO₂ asphyxiation, unless the animals were already dead. The spleen and a piece of liver (lower lobe) were aseptically removed, weighed, homogenized, serially diluted and played on LB agar and MacConkey agar for c.f.u. titres.

Pharmacokinetic analysis. Mice were injected intraperitoneally with a single dose of 50 mg kg⁻¹ darobactin, in 10% PEG-200. Blood samples were collected from *n* = 3 mice at each time point (0.25, 0.5, 1, 2, 3, 5, 8 and 24 h) via a tail snip, 10 µl of blood was diluted into 90 µl of chilled saline, and subsequently centrifuged at 1,000g for 5 min. The diluted plasma was decanted into a fresh tube and kept at –80 °C. Blood was collected from an untreated mouse and diluted in saline, and a standard curve generated by addition of known concentrations (0.1, 1, 10, and 100 µg ml⁻¹) of darobactin to decanted supernatant. All of the samples were run on LC-MS to determine the concentration of the compound in the blood. An Agilent 1260 Infinity liquid chromatography system and 6460 triple quadrupole (QQQ) system (Agilent Technologies) were used to quantify darobactin. A Thermo Scientific Accucore C18 column (50 mm × 2.1 mm, 2.6 µm) was used for the separation with a flow rate of 200 µl min⁻¹ with solvent A (0.1% (v/v) formic acid in Milli-Q water) and solvent B (0.1% (v/v) formic acid in acetonitrile). The initial concentration of 2% solvent B was maintained for 2 min, followed by a linear gradient to 70% over 10 min. MS parameters were as follows: gas temperature, 300 °C; gas flow, 7 l min⁻¹; capillary voltage, 3,500 V; fragmentor voltage, 100 V; scan type, MRM; transition parent ion 483.8 to product ions 211.3, 160.1, 120.1 and 103.1 with collision energy 42, 46, 50 and 94 V, respectively. MassHunter qualitative and quantitative analysis B.05 (Agilent Technologies) was used to quantify the darobactin peaks.

Septicaemia model. Darobactin was tested in a septicaemia protection model against *E. coli*, wild type (ATCC 25922) or multidrug-resistant (AR350, CDC), *P. aeruginosa*, wild-type PAO1 and a spontaneous polymyxin-resistant mutant (*pmrB* 523C>T mutation), and KPC *K. pneumoniae* (ATCC BAA1705). Mice were infected with 0.5 ml of bacterial

Article

suspension in BHI with 5% mucin (1×10^6 cells for *E. coli* and *K. pneumoniae*, 8×10^6 and 4×10^6 cells for *P. aeruginosa* wild-type and *pmrB* mutant strains, respectively) via intraperitoneal injection. This dose achieves >90% mortality within 24 h after infection. At 1 h after infection, mice received treatments with darobactin from 50 mg kg⁻¹ to 1 mg kg⁻¹ administered by intraperitoneal injection. Infection control mice were treated with 20 mg kg⁻¹ gentamicin as positive controls and the vehicle alone as a negative control. Survival was monitored for 7 days.

Thigh infection model. Darobactin was tested in a neutropenic thigh infection model against multidrug-resistant *E. coli* AR350 (CDC). Mice were rendered neutropenic through cyclophosphamide injections 4 days (150 mg kg⁻¹) and 1 day (100 mg kg⁻¹) before infection. An overnight culture (OD₆₀₀ of 2.0) of *E. coli* was diluted 1:1,000. Mice were infected with 100 µl of the prepared inoculum into the right thigh with the actual inoculum being 10^4 – 10^5 c.f.u. (determined by plate counts), and one group of mice was euthanized and the thighs collected and processed for time 0 counts. At 2 h after infection, mice received treatments with darobactin at 25 mg kg⁻¹ (given as 3 doses every 6 h) or 50 mg kg⁻¹ (given once) or gentamicin at 20 mg kg⁻¹ (one experiment, $n = 4$) or 50 mg kg⁻¹ (two experiments, $n = 5$ each) as a positive control, administered by intraperitoneal injection. At the time of treatment, one group of infected mice was euthanized, and thighs were collected and processed for c.f.u. At 26 h after infection, mice were euthanized via CO₂ asphyxiation. The right quadriceps muscles were aseptically removed, weighed, homogenized, serially diluted and plated on MHIIA for c.f.u. titres. This experiment was repeated on three separate occasions with one experiment containing $n = 4$ and two experiments containing $n = 5$ mice per group.

Statistics

Confidence intervals for IC₅₀ values in the BAM folding assay were calculated by GraphPad Prism v.8.2 using nonlinear regression [inhibitor] versus response, constraining the bottom to 0. Significance in transcriptome data for a differentially expressed gene was determined by $|\log_2(\text{fold change (FC)})| \geq 2$ and FDR-adjusted $P < 0.001$, differential expression was computed using the edgeR::exactTest⁴⁵ in R v.3.5.1 with unnormalized gene counts ($n = 4,626$ genes) for each treatment at time t versus matched control. For thermal proteome profiling, significant hits (FDR-adjusted $P < 1\%$) were calculated as previously described⁴⁹.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data supporting the findings of this study are available within the paper and its Supplementary Information or have been deposited to the indicated databases. The genome of *P. khanii* HGB1456 has been deposited to GenBank with accession number WHZZ000000000. The transcriptomic dataset (Extended Data Fig. 7) has been deposited to NCBI Sequence Read Archive with identifier PRJNA530781. The mass spectrometry proteomics (Extended Data Fig. 8 and Supplementary Table 2) data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD013319. Source Data for Figs. 2c, 4 and Extended Data Figs. 5b, 9 are provided with the paper. All other data are available from the corresponding author.

36. Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**, 1009–1015 (2016).
37. Blin, K. et al. antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* **45**, W36–W41 (2017).
38. Lassak, J., Henche, A. L., Binnenkade, L. & Thormann, K. M. ArcS, the cognate sensor kinase in an atypical Arc system of *Shewanella oneidensis* MR-1. *Appl. Environ. Microbiol.* **76**, 3263–3274 (2010).
39. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. USA* **97**, 6640–6645 (2000).
40. Tang, X. et al. Identification of thiotetronic acid antibiotic biosynthetic pathways by target-directed genome mining. *ACS Chem. Biol.* **10**, 2841–2849 (2015).
41. Gust, B., Challis, G. L., Fowler, K., Kieser, T. & Chater, K. F. PCR-targeted *Streptomyces* gene replacement identifies a protein domain needed for biosynthesis of the sesquiterpene soil odor geosmin. *Proc. Natl Acad. Sci. USA* **100**, 1541–1546 (2003).
42. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
43. Murphy, K. C. & Campellone, K. G. Lambda Red-mediated recombinogenic engineering of enterohemorrhagic and enteropathogenic *E. coli*. *BMC Mol. Biol.* **4**, 11 (2003).
44. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
45. Mateus, A. et al. Thermal proteome profiling in bacteria: probing protein state in vivo. *Mol. Syst. Biol.* **14**, e8242 (2018).
46. Becher, I. et al. Thermal profiling reveals phenylalanine hydroxylase as an off-target of panobinostat. *Nat. Chem. Biol.* **12**, 908–910 (2016).
47. Hughes, C. S. et al. Ultrasensitive proteome analysis using paramagnetic bead technology. *Mol. Syst. Biol.* **10**, 757 (2014).
48. Hughes, C. S. et al. Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat. Protoc.* **14**, 68–85 (2019).
49. Sridharan, S. et al. Proteome-wide solubility and thermal stability profiling reveals distinct regulatory roles for ATP. *Nat. Commun.* **10**, 1155 (2019).
50. Franken, H. et al. Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry. *Nat. Protoc.* **10**, 1567–1593 (2015).
51. Alvarez, F. J. D., Orelle, C. & Davidson, A. L. Functional reconstitution of an ABC transporter in nanodiscs for use in electron paramagnetic resonance spectroscopy. *J. Am. Chem. Soc.* **132**, 9513–9515 (2010).
52. Ritchie, T. K. et al. Chapter 11 - Reconstitution of membrane proteins in phospholipid bilayer nanodiscs. *Methods Enzymol.* **464**, 211–231 (2009).
53. Roman-Hernandez, G., Peterson, J. H. & Bernstein, H. D. Reconstitution of bacterial autotransporter assembly using purified components. *eLife* **3**, e04234 (2014).
54. Hagan, C. L., Kim, S. & Kahne, D. Reconstitution of outer membrane protein assembly from purified components. *Science* **328**, 890–892 (2010).

Acknowledgements This work was supported by NIH grant P01 AI118687 to K.L. and K.E.N. A. Mateus was supported by a fellowship from the EMBL Interdisciplinary Postdoc (EI3POD) Programme under Marie Skłodowska-Curie Actions COFUND (grant number 664726). S.H. was supported by the Swiss National Science Foundation via the NFP 72 (407240_167125). N.N. was supported by NIH grants GM127896 and GM127884. We thank H. Goodrich-Blair for providing strains of *Photobacterium* and *Xenorhabdus*; M. Kagan for help with isolating darobactin; the Northeastern University Barnett Institute MS Core Facility for access to its LC-MS resources; D. Baldisseri from Bruker Biospin Corporation for recording some of the NMR data of darobactin; N. Kurzwaga for the help with the analysis of thermal proteome profiling data; W. Fowle for assistance with scanning electron microscopy experiments; Y. Su for assistance with the ITC experiments; and R. Machado for help with taxonomy of *Photobacterium*.

Author contributions K.L. designed the study, analysed results and wrote the paper. Y.I. identified darobactin, designed the study and analysed results. K.J.M. designed the animal study, wrote the paper and analysed results. A.I. performed mass spectrometry and, with M.M., isolated darobactin. Q.F.-G., C.H., X.M., J.J.G. and A. Makriyannis identified the structure of darobactin. A.D. provided logistical support for the study. S.M. performed microscopy studies and analysed data. M.C. and M.G. performed susceptibility studies. S.N. performed animal studies. T.F.S., R.G., N.B., Z.G.W. and L.L.-O. identified darobactin BGCs and generated the knockout and heterologous expression strains. H.K. performed the NMR studies of BamA. S.H. designed and analysed the NMR studies and wrote the paper. R.W. performed the BAM nanodisc studies. N.N. designed and analysed the BAM nanodisc studies and wrote the paper. A.T., M.M.S. and A. Mateus performed the proteomics study and analysed data. K.E.N., J.L.E. and A.O. performed the transcriptome study and analysed data.

Competing interests The authors declare no competing interests.

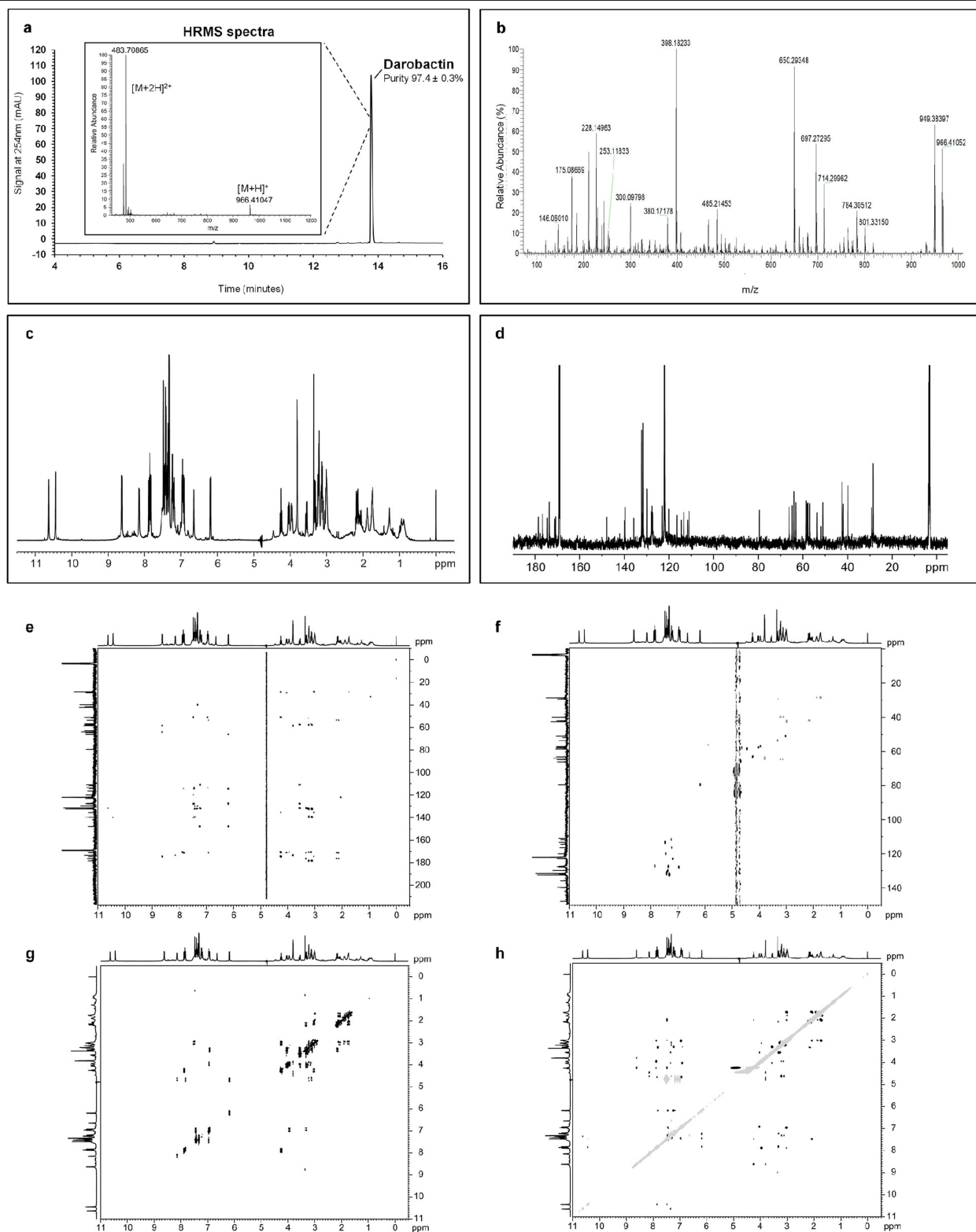
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1791-1>.

Correspondence and requests for materials should be addressed to K.L.

Peer review information Nature thanks Eric Brown, Tilmann Weber and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



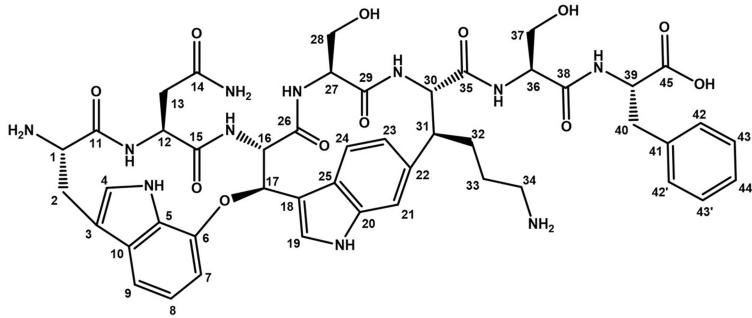
Extended Data Fig. 1 | Structural determination of darobactin. **a**, HPLC chromatogram of darobactin. Inset, high-resolution mass spectra (HRMS) of darobactin showing a peak at m/z 966.41047, which corresponds to the $[M+H]^+$ ion and another at m/z 483.70865, which corresponds to $[M+2H]^{2+}$ ion.

b, Higher-energy collisional dissociation-MS/MS spectra of darobactin. **c**, 1H NMR spectrum of darobactin. **d**, ^{13}C NMR spectrum. **e**, 1H - ^{13}C HMBC NMR spectrum. **f**, 1H - ^{13}C HSQC NMR spectrum. **g**, COSY NMR spectrum. **h**, ROESY NMR spectrum.

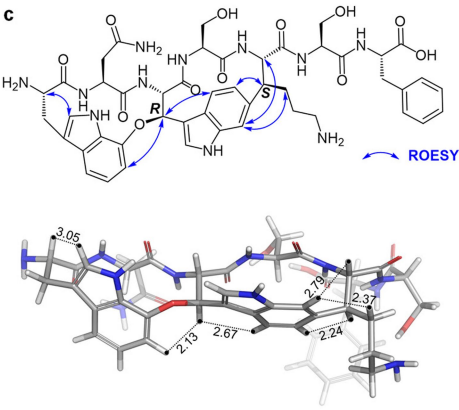
a

¹ H, ¹³ C and ¹⁵ N NMR chemical shifts (ppm) for Darobactin					
Position	δ _C / δ _N	δ _H (mult., J in Hz)	Position	δ _C / δ _N	δ _H (mult., J in Hz)
1	57.6	4.04 (1H, dd, 11.2, 7.7)	24	120.0	7.45 (1H, d, m)
1-NH ₂	-	exchanged	25	127.8	-
2	29.2	3.55 (1H, dd, 14.1, 7.6)	26	170.7	-
		3.30 (1H, dd, m)	26-NH	121.9	6.95 (1H, d, m)
3	111.0	-	27	56.9	3.95 (1H, m)
4	127.6	7.35 (1H, br s)	28	64.8	3.22 (1H, dd, m)
4-NH	128.8	10.63 (1H, br s)			3.14 (1H, dd, m)
5	131.8 [†]	-	29	170.9	-
6	147.9	-	29-NH	127.0	7.88 (1H, d, 10.7)
7	111.6	7.24 (1H, d, 7.7)	30	63.0	4.25 (1H, t, 10.9)
8	123.0	7.18 (1H, t, 7.7)	31	51.0	3.03 (1H, m)
9	116.5	7.22 (1H, d, 7.7)	32	28.5	2.08 (1H, m)
10	131.8 [†]	-	33	28.5	1.88 (1H, m)
11	171.1	-			1.74 (1H, m)
11-NH	124.5	6.92 (1H, d, 8.1)	34	42.4	2.99 (1H, m)
12	53.7	3.33 (1H, m)	34-NH ₂	-	7.51 (2H, v br s)
13	41.9	2.19 (1H, dd, 13.9, 7.2)	35	174.6	-
		2.13 (1H, dd, 13.9, 7.2)	35-NH	122.8	8.62 (1H, d, 7.3)
14	176.6	-	36	58.5	4.46 (1H, m)
14-NH ₂	113.8	7.31 (1H, br s)	37	64.0	3.80 (2H, d, 5.5)
		6.64 (1H, br s)	38	173.5	-
15	171.3	-	38-NH	123.6	8.14 (1H, d, 7.4)
15-NH	123.0	7.83 (1H, d, 9.8)	39	57.9	4.64 (1H, dt, 7.7, 5.9) [†]
16	66.1	4.69 (1H, dd, 9.1, 10.2) [†]	40	39.8	3.11 (1H, dd, m)
17	79.5	6.18 (1H, d, 9.1)			3.22 (1H,dd, m)
18	114.5	-	41	139.6	-
19	127.4	7.85 (1H, br s)	42, 42'	132.3	7.32 (2H, d, m)
20	139.9	-	43, 43'	131.5	7.42 (2H, t, 7.49)
20-NH	133.2	10.44 (1H, br s)	44	129.9	7.37 (1H, t, 7.08)
21	113.3	7.48 (1H, br s)	45	178.4	-
22	135.7	-			
23	127.7	6.96 (1H, d, m)			

b

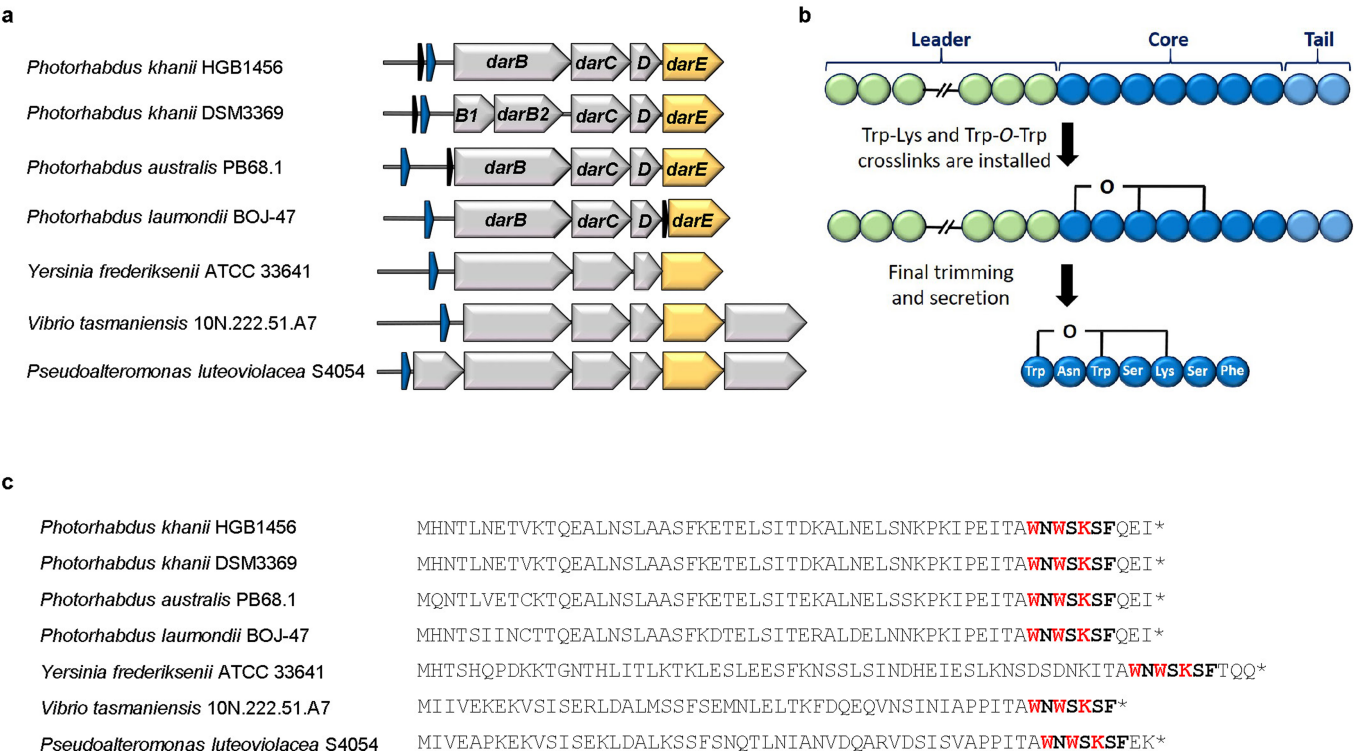


c



Extended Data Fig. 2 | NMR assignments of darobactin. **a**, ¹H, ¹³C and ¹⁵N NMR chemical shifts (ppm) for darobactin. [†]Owing to overlap with a residual water peak at 4.6 ppm, the multiplicity and J coupling values were from a different ¹H NMR spectrum of darobactin in water:deuterated acetonitrile (2:1, v/v). [‡]Two

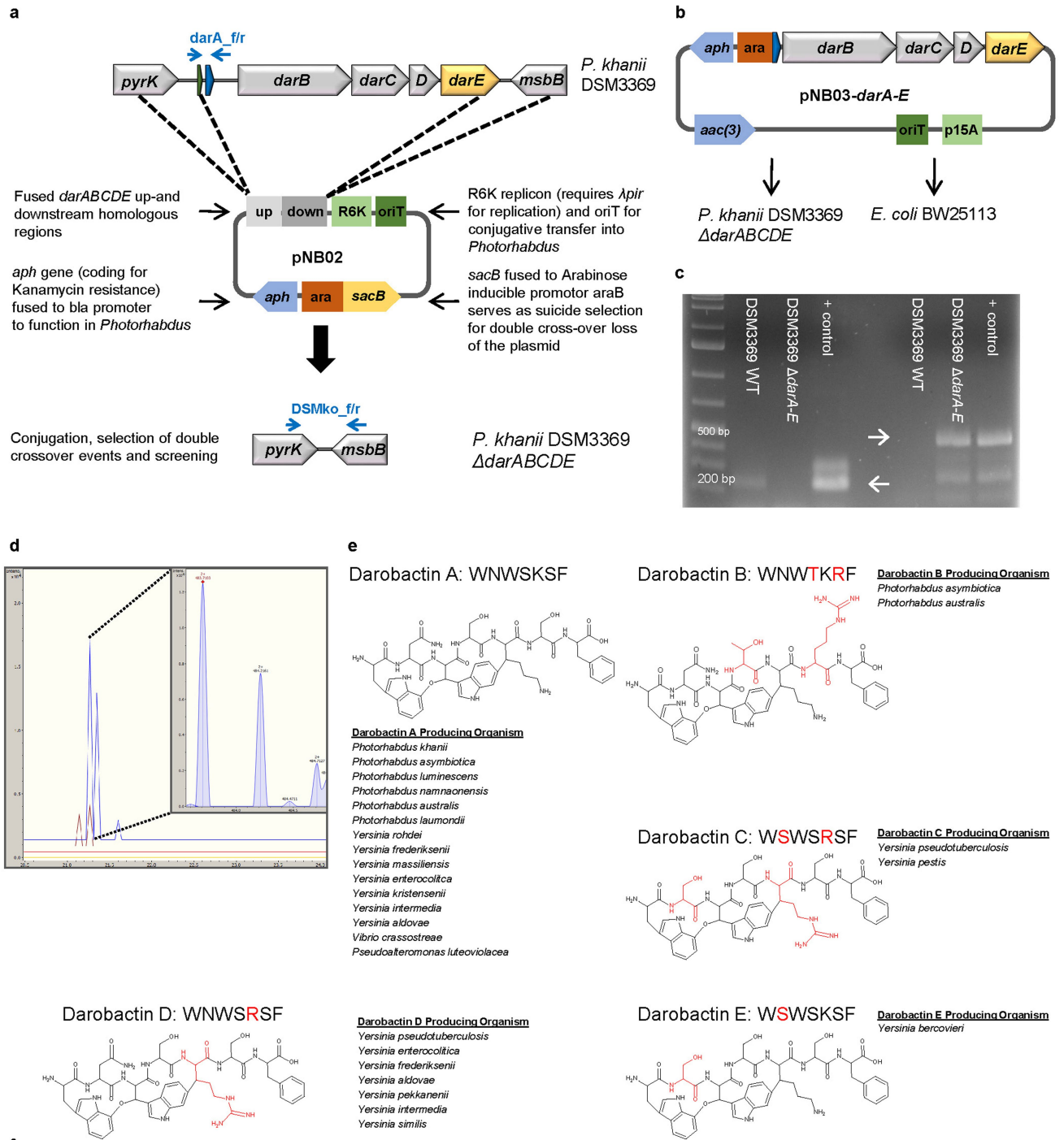
partially overlapping peaks were observed at 131.79 ppm and 131.83 ppm. **b**, Structure of darobactin with numbering for NMR assignments. **c**, Key ROESY correlations (top) and three-dimensional model of darobactin (bottom).



Extended Data Fig. 3 | BGC of darobactin in selected bacterial strains.

a, The BGC consists of the structural gene *darA* (coloured in blue), *darBCD* (transporter encoding genes; grey) and *darE* (a radical SAM enzyme; orange). In addition, a *relE*-like gene (black) open-reading frame is co-located with the BGC at different positions in different species. The BGC can be detected in most *Photorhabdus* strains in a conserved genetic region. In addition, homologous BGCs (related genes show the identical colour code) can be found in *Yersinia*,

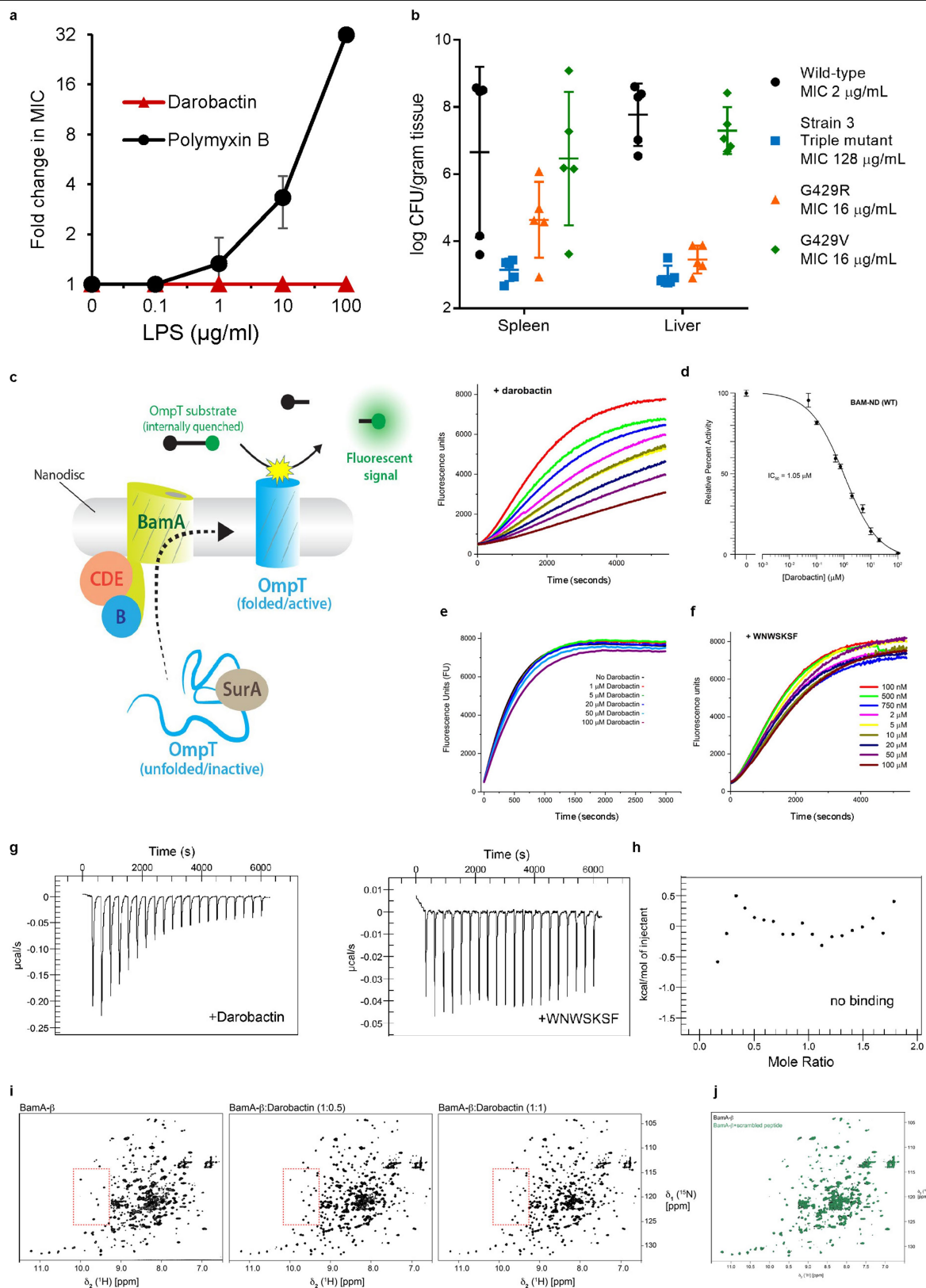
Vibrio and *Pseudoalteromonas* strains. **b**, Biosynthetic hypothesis. The propeptide encoded by *darA* consists of 58 amino acids. The crosslinks are installed on the linear propeptide by DarE. In a next step, the leader and tail regions are cleaved off and darobactin is secreted by the ABC transporter DarBCD. **c**, The amino acid sequence of the propeptide from selected bacterial strains. The darobactin core peptide is highlighted in bold and the amino acids involved in the crosslinking in bold red. The asterisk indicates the stop codon.



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Darobactin knockout strain and heterologous expression, and putative structures and producers of darobactin A–E. **a**, Schematic of the double crossover knockout vector pNB02 and the targeted genomic region. **b**, Schematic of the darobactin BGC expression plasmid. **c**, Test PCRs on *P. khanii* DSM3369 $\Delta darABCDE$, showing the loss of the darobactin BGC. Left, amplification of *darA* (primers darA_f/r) results in a 177-bp fragment in the wild-type (WT) strain and in no fragment in the mutant. Right, after loss of pNB02 (indicated by sensitivity to kanamycin), amplification of a 450-bp fragment if the BGC is deleted (primers DSMko_f/r) occurs. Positive controls include pNB03-*darA-E* and pNB02. Primer positions are indicated in blue in **a**. The raw DNA gel is provided in Supplementary Fig. 1. **d**, LC-MS-extracted ion chromatogram at $m/z = 483.7089 \pm 0.001$. Yellow,

P. khanii DSM3369 $\Delta darABCDE$ and pNB03; red, *P. khanii* DSM3369 $\Delta darABCDE$ and pNB03-*darA-E*; brown, *E. coli* BW25113 and pNB03-*darA-E*; blue, *P. khanii* DSM3369 wild type. Inset, HRMS spectrum of the ion peak showing the double charged $[M + 2H]^{2+}$ ion that corresponds to darobactin. **c, d**, Data are representative of at least three independent biological replicates. **e**, Putative darobactin analogues B–E were drawn based on the amino acid sequence that is present in the darobactin BGC. The proposed darobactin-producing organisms were identified by a BLASTp search of the seven-amino-acid sequence of darobactin A, confirming the presence of *darBCDE* downstream of the propeptide. Amino acid changes from darobactin A are highlighted in red. **f**, The propeptide sequence of the various darobactin analogues.

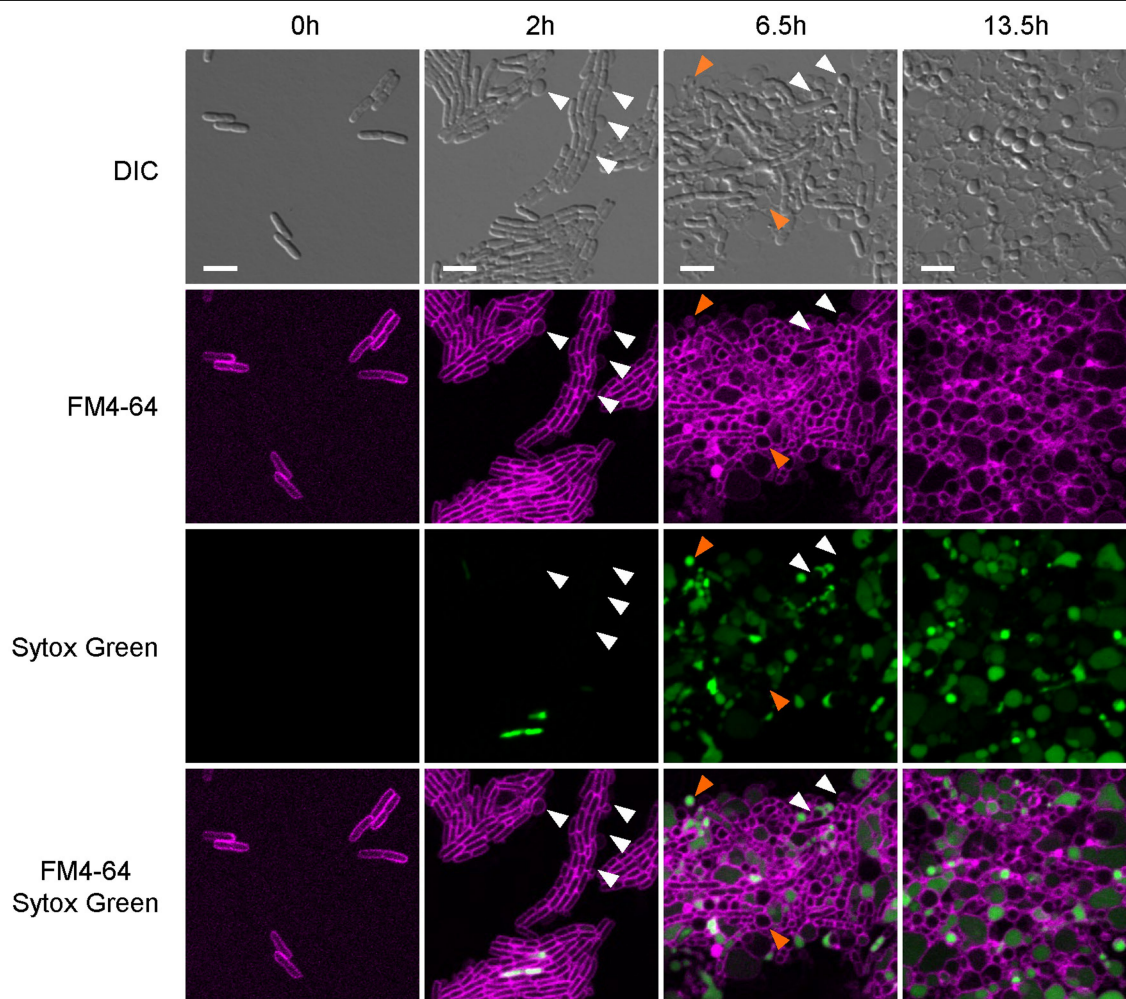


Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Darobactin mechanism of action and resistance

studies. a, Darobactin and polymyxin B MIC studies against *E. coli* MG1655 were performed in the presence of LPS. Addition of LPS antagonized polymyxin activity, but not darobactin activity. Data are mean \pm s.d. of triplicate experiments. **b,** Groups of five mice were infected intraperitoneally with 10^7 *E. coli* ATCC 25922, and subsequently euthanized at 24 h (if not already dead), after which the livers and spleens collected, homogenized and plated for c.f.u. analysis. Wild-type *E. coli* caused 60% death and showed high c.f.u. burdens in liver and spleen. All three darobactin-resistant *bamA* mutant strains had reduced virulence, with 100% survival in all groups at 24 h. The burden of bacteria of the triple *bamA* mutant was close to the limit of detection in organs, the G429R-expressing mutant was found at low but detectable levels, whereas the G429V-expressing mutant was found at relatively high loads in the organs. $n = 5$. Data are mean \pm s.d. **c,** Schematic of the BAM activity assay in which BAM (BamA-E) was first inserted into lipid nanodiscs. Unfolded OmpT, along with the periplasmic chaperone SurA, was then mixed with the BAM-nanodiscs, and BAM folds OmpT into the nanodisc. OmpT, a protease, cleaves an internally quenched peptide, which produces a fluorescent signal. **d,** BAM-nanodisc assays performed in the presence of increasing concentrations of darobactin (left). The results show that darobactin is able to specifically inhibit BAM-nanodisc activity in a dose-dependent manner. These data were then normalized against the 'no darobactin' sample and the highest concentration of darobactin and plotted, and an IC_{50} was calculated using the online IC_{50} calculator tool (AAT Bioquest) (right). ND, nanodisc. $n = 3$ biologically independent experiments. Data are mean \pm s.d. **e,** As a control to the BAM-nanodisc assays, we prepared OmpT-nanodiscs and assayed OmpT-nanodisc activity in the presence of increasing concentrations of darobactin. To prepare

the OmpT-nanodiscs, we first expressed OmpT as inclusion bodies and then refolded the protein using previously reported methods^{53,54}. We then incorporated OmpT into nanodiscs using the same methods as described for BAM. The assays were performed using 0.4 μ M of OmpT-nanodiscs. The results show that darobactin has almost no effect on OmpT-nanodisc activity, thus confirming that darobactin does not affect OmpT activity itself or disrupting the nanodiscs themselves. A representative plot is shown from a triplicate experiment. **f,** The WNWSKSF peptide does not inhibit BAM-nanodiscs. As a control to darobactin, the BAM-nanodisc assays were performed in the presence of increasing concentrations of a linear peptide WNWSKSF. The results show that the WNWSKSF peptide has only minimal effects on BAM-nanodisc activity, even at the highest concentrations. A representative plot is shown from a triplicate experiment. **g, h,** Specific binding of darobactin to BamA/BAM. Mole ratio is the protein:ligand ratio. **g,** Plot of ITC experiments of wild-type BAM titrated with darobactin. $K_d = 1.2 \mu$ M, $N = 0.52$, $\Delta H = -25 \text{ kcal mol}^{-1}$ and $\Delta S = -56 \text{ cal mol}^{-1} \text{ K}^{-1}$. The experiment was repeated independently twice with similar results. **h,** Plot of ITC experiments of wild-type BAM titrated with the peptide WNWSKSF shows that there is no binding within the same concentration range as was used for darobactin. The experiment was repeated independently twice with similar results. **i, j,** Two-dimensional [¹⁵N, ¹H]-TROSY spectra of 250 μ M BamA- β in 0.1% w/v LDAO. **i,** BamA- β in the absence (left) and in the presence of darobactin with a molar ratio of 1:0.5 (middle) and 1:1 (right) of BamA- β :darobactin. The red dashed line outlines an example spectral region that shows substantial spectral changes during the titration. The experiment was repeated independently twice with similar results. **j,** An overlay of apo BamA- β (black) (250 μ M) on BamA- β and a scrambled linear peptide WNKWSFS (green) (230 μ M). The experiment was performed once.

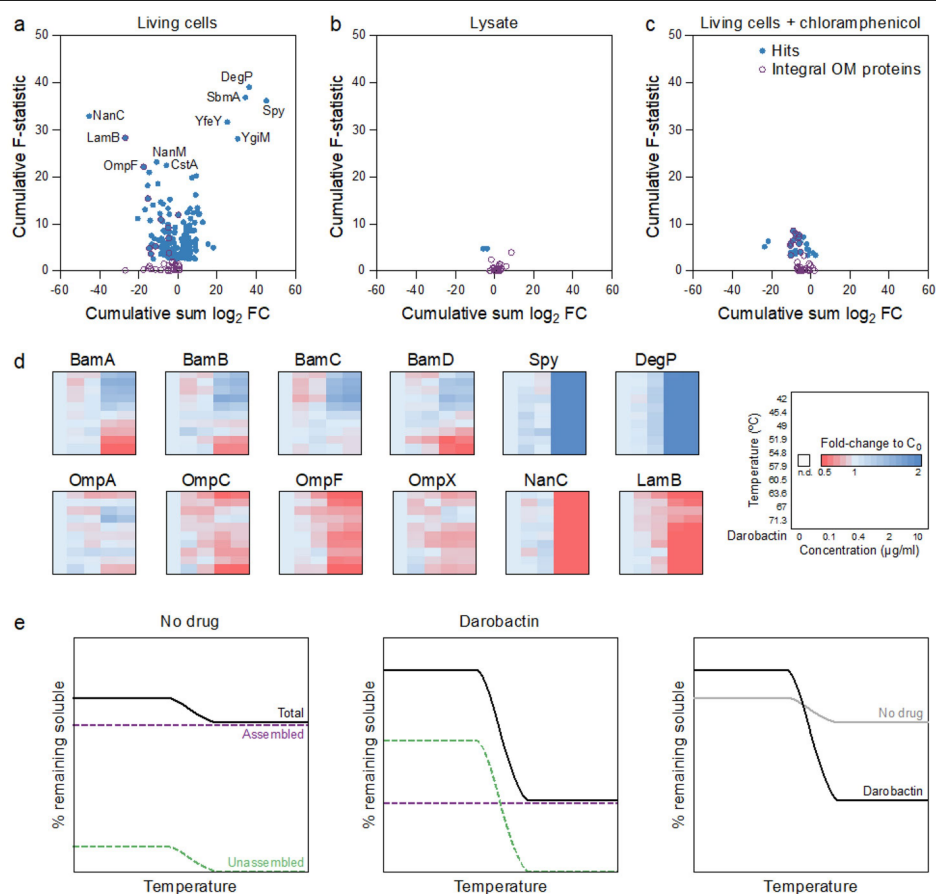


Extended Data Fig. 6 | Darobactin disrupts the outer membrane and causes lysis of *E. coli*. *E. coli* MG1655 cells were placed on top of an agarose pad that contained darobactin and the fluorescent dyes FM4-64—to stain the membrane (false-coloured in magenta)—and Sytox Green—to show membrane permeabilization (false-coloured in green). *E. coli* MG1655 cells were observed

over time at 37 °C under the microscope. For each indicated time point, representative panels show the killing progression of *E. coli* MG1655 with darobactin. White arrows highlight membrane blebbing; orange arrows highlight swelling and lysis. Scale bars, 5 μm. This figure is representative of three biologically independent experiments performed with similar results.

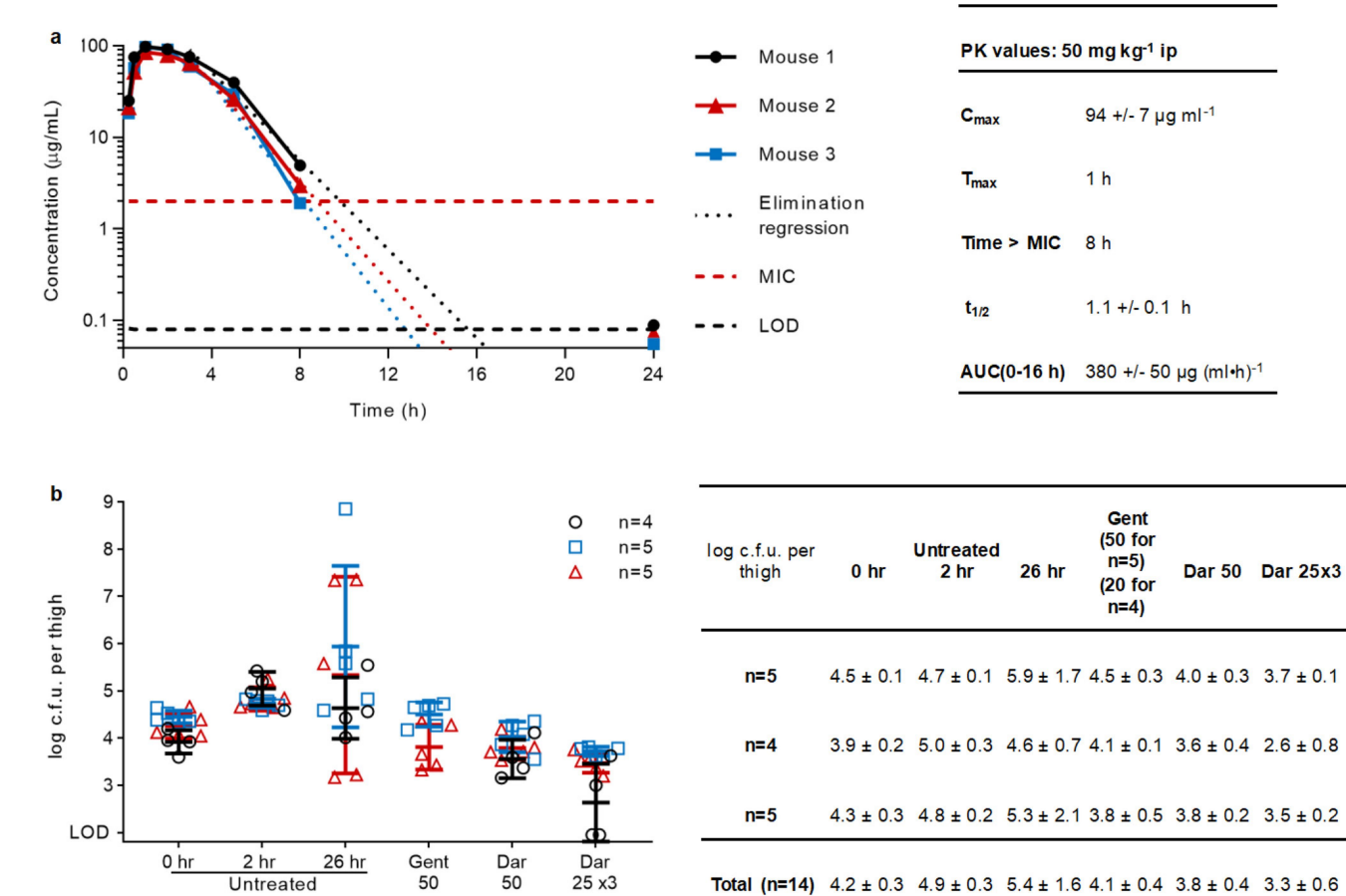
Extended Data Fig. 7 | Transcriptome analysis of darobactin treatment shows activation of envelope stress pathways. *E. coli* BW25113 were treated with 1× MIC darobactin, and the RNA isolated and sequenced. **a–c**, Volcano plots illustrate differential gene expression (Fisher’s exact test in edgeR; results were deemed significant if $|\log_2(\text{FC})| \geq 2$ and FDR-corrected $P < 0.001$; $n = 3$ biologically independent samples for each control or treatment sample) at time points $t = 15$ min (**a**), $t = 30$ min (**b**) and $t = 60$ min (**c**) after exposure. Grey region, not significant. **d**, Network visualization of differentially expressed genes at each time point. Nodes include genes (coloured circles) and time

points (grey rectangles). Gene node colours represent relevant functional categories. Directed edges radiating from a time point node represent differentially expressed genes with respect to the given time point with weights reflecting the $|\log_2(\text{FC})|$. **e**, Right, heat map showing the differential expression ($|\log_2(\text{FC})|$) of genes of interest. Left, assignment to envelope stress pathways. Solid lines depict members of the same operon. In all panels, red indicates downregulation (lower expression in treatment relative to control) and blue indicates upregulation.



Extended Data Fig. 8 | Two-dimensional thermal proteome profiling of darobactin. **a–c**, Pseudo-volcano plots for two-dimensional thermal proteome profiling experiments of darobactin treatment (10 min) of *E. coli* BW25113 in living cells (**a**), lysates (**b**) and living cells pre-treated with chloramphenicol to inhibit protein synthesis (**c**). $n = 1$ for each concentration, heated to 10 different temperatures, for each experiment. Significant hits (FDR-adjusted $P < 1\%$, calculated with a functional analysis of dose–response, requiring stabilization

effects at $n > 1$ temperatures as described previously⁴⁹) are highlighted in blue and integral outer membrane proteins are highlighted in purple. **d**, Heat maps for selected proteins in the experiment with living cells. For each protein and temperature (a key is shown on the right), the signal intensity was normalized to the vehicle control. **e**, Schematic of putative thermally stable assembled versus labile unassembled populations of the BAM machinery with darobactin treatment.



Extended Data Fig. 9 | Darobactin single-dose pharmacokinetics and mouse thigh models. **a**, Three mice were intraperitoneally injected with 50 mg kg⁻¹ darobactin, and blood samples were collected by tail snip over 24 h. Samples ($n=1$ per time point and mouse) were analysed for darobactin content by LC-MS/MS, and concentrations were calculated using a standard curve created by linear regression on the log(area under the curve peak) to log(concentration) of standards. Pharmacokinetic values were calculated in Excel; $t_{1/2}$ and time > MIC assuming first-order elimination and using linear regression on

time points 3–8 h; AUC (0–16 h) using the trapezoid rule. The limit of detection (LOD) was 0.08 $\mu\text{g mL}^{-1}$. **b**, A mouse thigh model was repeated three times testing the efficacy of darobactin against *E. coli* AR350. Mice were injected with bacteria in their right thigh at 0 h, then dosed with no drug, gentamicin or darobactin starting 2 h after infection (50 mg kg⁻¹ once, 25 mg kg⁻¹ given three times every 6 h, or 20 mg kg⁻¹ once). At 26 h mice were euthanized and thighs collected and homogenized tissues were plated for c.f.u. analysis. Data are mean \pm s.d.

Extended Data Table 1 | *Photorhabdus* and *Xenorhabdus* species

<i>Photorhabdus</i> sp.	# Strains in screen	Source	<i>Xenorhabdus</i> sp.	# Strains in screen	Source
<i>P. akhurstii</i>	1	DSMZ*	<i>X. beddingii</i>	1	HGB
<i>P. caribbeanensis</i>	1	DSMZ	<i>X. bovienii</i>	12	HGB
<i>P. cinerea</i>	1	DSMZ	<i>X. doucetiae</i>	1	HGB
<i>P. hainanensis</i>	1	DSMZ	<i>X. indica</i>	5	DSMZ
<i>P. heterorhabditis</i>	2	DSMZ	<i>X. innexi</i>	3	HGB and DSMZ
<i>P. kayaii</i>	1	DSMZ	<i>X. ishibashi</i>	1	DMZ
<i>P. khanii</i>	2	HGB† and DSMZ	<i>X. japonica</i>	2	HGB and DSMZ
<i>P. kleinii</i>	1	DSMZ	<i>X. japonicus</i>	1	HGB
<i>P. laumondii</i> subsp. <i>laumondii</i>	1	DSMZ	<i>X. khoisanae</i>	4	DSMZ
<i>P. luminescens</i>	3	HGB	<i>X. miraniensis</i>	2	HGB
<i>P. noenieputensis</i>	1	DSMZ	<i>X. nematophila</i>	2	HGB
<i>P. stackebrandtii</i>	1	DSMZ	<i>X. poinarii</i>	3	HGB
<i>P. tasmaniensis</i>	1	DSMZ	<i>X. szentirmaii</i>	1	HGB
<i>P. temperata</i>	6	HGB and DSMZ			
<i>P. thracensis</i>	1	DSMZ			
<i>Photorhabdus</i> sp.	5	HGB			

*DSMZ, Deutsche Sammlung von Mikroorganismen und Zellkulturen.

†HGB, H. Goodrich-Blair.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☐ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Xcalibur software (Thermo Fisher Scientific Inc.) was used for mass spectrometry analysis. Schrodinger 2018-2 was used for molecular modeling. SPAdes 3.11 was used to assemble the *Photobacterium* genome. antiSMASH v4 was used to analyze the genome for BGCs. BLAST was used to search for homologous operons. Zen Software v14.03.201 was used to acquire fluorescent microscopy images. IsobarQuant and Mascot 2.4 were used to generate proteomic data. MassHunter B0.5 was used to quantify darobactin peaks in quantitative mass spectrometry experiments.

Data analysis

Microsoft Excel and GraphPad Prism 8.2 were used to plot graphs. Geneious 11.0.4 was used to find mutations in the resistant mutants. Fiji v1.52i software was used to process microscopy images. R v3.5.1 was used to analyze the transcriptomic and proteomic data, and Python v3.6.6 was used to plot transcriptomic data. AAT Bioquest was used to plot the BAM folding assay ITC data was analyzed and fit to the independent binding model using the NanoAnalyze software package.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The genome of *P. kharii* HGB1456 has been deposited to Genbank with identifier WHZZ000000000.

The transcriptomic dataset (Extended Data Figure 7) has been deposited to NCBI Sequence Read Archive with identifier PRJNA530781.

The proteomics (Extended Data Figure 8) data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with

the dataset identifier PXD013319.

All other data available from corresponding author on reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For animal studies, sample size was chosen based on prior experience with variability between mice for infection models. For survival analysis in septicemia, three mice were used as difference in survival with treatment is dramatic. For the thigh infection model, five mice per group was used due to known variability in infection burdens in untreated mice. No sample size calculation was performed.
Data exclusions	No data was excluded
Replication	For standard microbiological assays, MICs and time-kill, experiments were done in triplicate to ensure findings were reproducible. Resistant mutants were also generated from three independent cultures. In animal studies, multiple mice were included in each group.
Randomization	Mice were randomly assigned to groups for infection and treatment, and were then housed as a group in a cage for the duration of the experiment.
Blinding	Blinding was deemed not relevant to mouse work, as outcomes were quantitative and not subjective.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	All cell lines were obtained from ATCC or GenTarget in 2018
Authentication	None of the cell lines were authenticated
Mycoplasma contamination	The cell lines were not tested for mycoplasma contamination
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Female CD-1 mice, 6 weeks old were used for all animal studies
Wild animals	The study did not involve wild animals

Field-collected samples

Photorhabdus and Xenorhabdus strains were received from Heidi Goodrich-Blair, grown at 28 C on LB agar, then stored at -80 C in glycerol stocks.

Ethics oversight

Animal studies were approved by the Northeastern University IACUC.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

An intra-tumoral niche maintains and differentiates stem-like CD8 T cells

<https://doi.org/10.1038/s41586-019-1836-5>

Received: 11 December 2018

Accepted: 13 November 2019

Published online: 11 December 2019

Caroline S. Jansen¹, Nataliya Prokhnevskaya¹, Viraj A. Master^{1,2}, Martin G. Sanda^{1,2}, Jennifer W. Carlisle^{2,3}, Mehmet Asim Bilen^{2,3}, Maria Cardenas¹, Scott Wilkinson⁴, Ross Lake⁴, Adam G. Sowalsky⁴, Rajesh M. Valanparambil^{5,6}, William H. Hudson^{5,6}, Donald McGuire^{5,6}, Kevin Melnick¹, Amir I. Khan¹, Kyu Kim¹, Yun Min Chang⁵, Alice Kim¹, Christopher P. Filson^{1,2}, Mehrdad Alemozaffar^{1,2}, Adeboye O. Osunkoya^{1,2,7}, Patrick Mullane⁷, Carla Ellis⁷, Rama Akondy^{5,6}, Se Jin Im^{5,6}, Alice O. Kamphorst⁸, Adriana Reyes¹, Yuan Liu^{2,9} & Haydn Kissick^{1,2,5,6*}

Tumour-infiltrating lymphocytes are associated with a survival benefit in several tumour types and with the response to immunotherapy^{1–8}. However, the reason some tumours have high CD8 T cell infiltration while others do not remains unclear. Here we investigate the requirements for maintaining a CD8 T cell response against human cancer. We find that CD8 T cells within tumours consist of distinct populations of terminally differentiated and stem-like cells. On proliferation, stem-like CD8 T cells give rise to more terminally differentiated, effector-molecule-expressing daughter cells. For many T cells to infiltrate the tumour, it is critical that this effector differentiation process occur. In addition, we show that these stem-like T cells reside in dense antigen-presenting-cell niches within the tumour, and that tumours that fail to form these structures are not extensively infiltrated by T cells. Patients with progressive disease lack these immune niches, suggesting that niche breakdown may be a key mechanism of immune escape.

In many cancers, tumour-infiltrating CD8 T cells predict patient survival and response to immunotherapy^{1–8}. These observations raise a fundamental question about the immune response to cancer and why some tumours have high CD8 T cell infiltration while others do not. A logical assumption has been made that T cell exhaustion drives a decline in the T cell response. T cell exhaustion has been extensively described in viral infections, in which persistent antigen exposure reduces the ability of the CD8 T cells to proliferate and kill target cells^{9,10}. Acquisition of checkpoint molecules that inhibit T cell function are a hallmark of this exhausted state, and blockade of molecules such as PD-1 can rescue exhausted cells in these models^{11,12}. Supporting the idea that T cell exhaustion is a factor that limits T cell function in cancer, many reports have found that T cells in tumours express high levels of these checkpoint molecules, and blockade of PD-1 and CTLA-4 are among the most successful treatments for many cancers^{13–17}. However, the model of persistent antigen exposure driving T cell decline does not explain why some patients have a strong T cell response to their tumour for decades, or why patients with controlled disease may have many CD8 T cells that are phenotypically exhausted. Here we investigate the CD8 T cell response to human tumours to better explain the mechanisms that control the magnitude of the T cell response to cancer.

TCF1⁺ CD8 T cells reside in tumours

On the basis of the observation that CD8 infiltration into tumours predicts survival and response to immunotherapy in other cancers^{1–7,18,19}, we

measured this parameter in a cohort of patients with kidney cancer. To quantitate CD8 infiltration, tumour tissue was collected from patients undergoing surgery and analysed by flow cytometry (Extended Data Fig. 1a). CD8 T cell infiltration ranged from 0.002% to over 20% of the total tumour cells (Fig. 1a). For patients with disease at any stage, having less than 2.2% CD8 T cell infiltration predicted four-fold more rapid progression after surgery (hazard ratio (HR) = 3.84, $P < 0.01$) (Fig. 1b, Extended Data Figs 1b–e, 2a, b). CD8 T cell infiltration did not correlate with clinical parameters such as disease stage or patient age (Extended Data Fig. 2c–k), suggesting that other biological mechanisms control the degree of T cell infiltration into tumours.

Reasoning that the composition of the tumour-infiltrating CD8 T cells might offer insight into the mechanisms controlling cell infiltration, we analysed expression of checkpoint molecules, co-stimulatory molecules and important transcription factors in tumour-infiltrating CD8 T cells. We detected a distinct population of cells that resembled exhausted CD8 cells by their expression of high levels of checkpoint molecules, TIM3, PD-1, CTLA4 and TIGIT (Fig. 1c, d, Extended Data Fig. 3a, b). We also identified a population of cells with low checkpoint molecule expression, but high expression of co-stimulatory molecule CD28 and transcription factor TCF1 (encoded by *TCF7*) (Fig. 1c, d, Extended Data Fig. 3a, b). TCF1 is a critical transcription factor that defines a stem-like T cell population in chronic murine lymphocytic choriomeningitis virus (LCMV) infection^{20–22}. Of note, others have described a TCF1⁺ CD8 T cell population in human

¹Department of Urology, Emory University School of Medicine, Atlanta, GA, USA. ²Winship Cancer Institute of Emory University, Atlanta, GA, USA. ³Department of Hematology and Oncology, Emory University School of Medicine, Atlanta, GA, USA. ⁴Laboratory of Genitourinary Cancer Pathogenesis, National Cancer Institute, Bethesda, MD, USA. ⁵Department of Microbiology and Immunology, Emory University School of Medicine, Atlanta, GA, USA. ⁶Emory Vaccine Centre, Emory University School of Medicine, Atlanta, GA, USA. ⁷Department of Pathology, Emory University School of Medicine, Atlanta, GA, USA. ⁸Department of Oncological Sciences, Precision Immunology Institute, Icahn School of Medicine at Mount Sinai, New York City, NY, USA. ⁹Rollins School of Public Health, Emory University, Atlanta, GA, USA. *e-mail: haydn.kissick@emory.edu

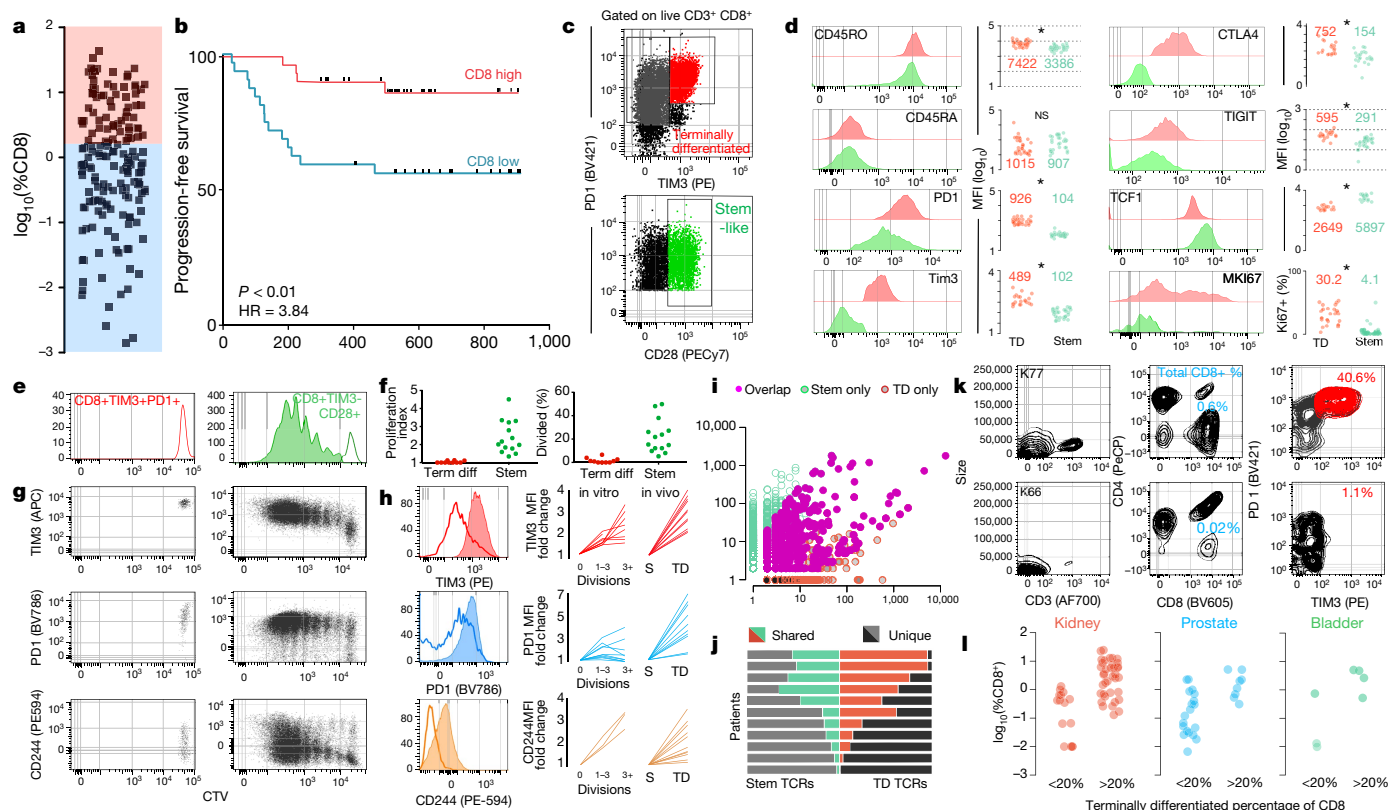


Fig. 1 | The anti-tumour T cell response is supported by a stem-like CD8 T cell, which gives rise to terminally differentiated CD8 T cells in the tumour.

a, Proportion of CD8 T cells in kidney tumours shown as percent of total cells ($n = 68$). **b**, Disease progression after surgery in patients with kidney cancer stratified into high or low CD8 T cell infiltration ($\pm 2.2\%$) based on optimal cut methods. Time to progression is the number of days from surgery until death or progression by RECIST criteria ($n = 66$). **c**, Gating strategy to identify intra-tumoral CD8 T cell populations. Populations shown are gated on live, CD3⁺ and CD8⁺. **d**, Expression (mean fluorescence intensity (MFI)) of activation markers, checkpoint molecules and transcription factors by TIM3⁺ and TIM3⁻ CD28⁺ subsets, gated as in **c**. **e, f**, Stem-like (TIM3⁻ CD28⁺) and terminally differentiated (TIM3⁺) populations were sorted from kidney tumours, labelled with CellTrace violet, and cultured with anti-CD3/anti-CD28 beads and 10 U ml⁻¹ of IL-2 for 4–5

and mouse tumours that correlates with response to PD-1 blockade^{22–27}. To functionally characterize the TCF1⁺ and checkpoint-high populations of CD8 T cells in tumours, checkpoint-high cells (PD-1⁺, TIM3⁺) and stem-like cells (TCF1⁺ TIM3⁻ CD28⁺) were sorted from tumours, labelled with CellTrace Violet, and incubated with anti-CD3/CD28 stimulation beads. The TCF1⁺ TIM3⁻ CD28⁺ stem-like population consistently proliferated in response to bead stimulus, whereas the checkpoint-high population lacked proliferative potential (Fig. 1e, f). Of note, after division, the stem-like T cells upregulated PD-1, TIM3 and CD244 to a similar level seen in vivo and downregulated TCF1, acquiring the phenotype of the checkpoint-high population (Fig. 1g, h, Extended Data Fig. 3c–f). Together, these data suggest that TIM3⁻ CD28⁺ T cells possess a stem-like capability; they can proliferate and give rise to more terminally differentiated, checkpoint-molecule-expressing T cells.

To further investigate the relationship between the intra-tumoral stem-like and terminally differentiated CD8 T cells, we examined the T cell receptor (TCR) repertoires of each population in 11 tumour samples. We found that TCRs significantly overlapped between the stem-like and terminally differentiated cell populations in all patients examined, suggesting a clonal relationship between these populations (Fig. 1i, j, Extended Data Fig. 4h). In two patients from whom we

days. Proliferation index and percentage of cells divided is shown.

g, h, Expression of TIM3, PD-1 and CD244 after cells undergo proliferation. Summary plots from in vitro activation experiments compared to fold change in MFI observed between the populations in vivo. **i**, TCR repertoires of stem-like and terminally differentiated T cells sorted as shown in Extended Data Fig. 4. TCR clones are represented by the number of reads detected in either T cell population. **j**, TCR repertoire overlap between stem-like and terminally differentiated T cells. The proportion of the detected TCR repertoire in each patient that is unique to each population or shared between the two is shown. **k, l**, Generation of checkpoint-high cells correlates with total T cell infiltration. Patients were classified as having a low (<20%) or high (>20%) fraction of TIM3⁺ terminally differentiated cells. Data show sample patients (**k**), and summary data in kidney ($n = 49$), prostate ($n = 28$) and bladder tumours ($n = 8$) (**l**).

recovered samples from distant sites within the same tumour, we found a high degree of TCR overlap between the stem-like and terminally differentiated populations at all locations (Extended Data Fig. 4g). These data are in contrast to reports finding that the CD39⁺ population of tumour-infiltrating lymphocytes (TILs) are unrelated to tumour antigens, and instead support a model of T cell differentiation whereby stem-like T cells within the tumour are the precursors to the terminally differentiated CD8 T cell population²⁸.

We next assessed how the composition of CD8 T cells in the tumour related to total T cell infiltration. Highly infiltrated tumours consistently had a distinct population of TIM3⁺ cells, which resemble phenotypically exhausted CD8 T cells, whereas poorly infiltrated tumours rarely had these cells (Fig. 1k, Extended Data Fig. 3g). The same relationship was evident in prostate and bladder tumours as well, where poorly infiltrated tumours contained few TIM3⁺ terminally differentiated cells (Fig. 1l). In poorly infiltrated tumours, the stem-like CD8 T cell population is consistently detectable at very low numbers (Extended Data Fig. 3h) but does not appear to be induced to differentiate into the TIM3⁺ cells (Fig. 1k, l, Extended Data Fig. 3g). These data suggest that the magnitude of the T cell response within a tumour is related to the ability of many terminally differentiated cells to be generated by the stem-like TCF1⁺ T cell population.

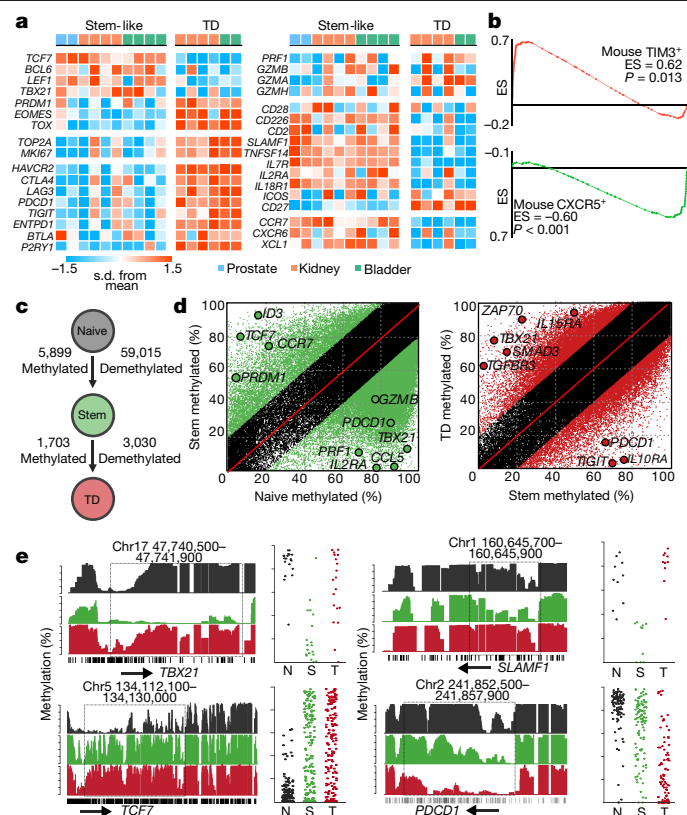


Fig. 2 | Stem-cell differentiation to the terminally differentiated state is associated with transcriptional and epigenetic changes. **a**, Heat map of transcription factors, proliferation-related genes, checkpoint molecules, cytotoxic molecules, co-stimulatory molecules, survival genes and migration and adhesion genes. Figure shows the z-scored data. TD, terminally differentiated. **b**, GSEA comparison to mouse CXCR5⁺ and TIM3⁺ subsets of CD8 T cells. Gene sets were created from CXCR5 stem-like and TIM3⁺ exhausted CD8 subsets from LCMV infection. Plots show enrichment score (ES) against genes upregulated (red) and downregulated (green) in mice. **c**, Summary of the number of epigenetic changes occurring as CD8 T cells undergo differentiation. Illustration shows the number of DNA methylation changes occurring as cells differentiate. **d**, Green regions show methylated and demethylated regions as cells transition from naive to stem-like cells, and red shows these events as cells transition from stem-like to terminally differentiated. **e**, Specific epigenetic changes near important differentially expressed genes. Histograms show the total methylation from 0–100% in regions near important genes. Highlights show significantly differentially methylated regions. Dot plots show the methylation of each CpG motif within this highlighted domain.

Transcription and epigenetics of CD8 T cell subsets

To further investigate the terminally differentiated and stem-like T cell populations in tumours, we performed RNA-sequencing (RNA-seq) on these populations. The terminally differentiated cells expressed more checkpoint molecules and much higher levels of granzymes and perforin (Fig. 2a). In contrast, TCF1⁺ stem-like CD8 T cells had higher levels of genes involved in survival such as *IL7R* and *IL2RA* (CD25), as well as co-stimulatory molecules such as *CD28*, *CD226* and *CD2* (Fig. 2a). We also compared these populations to stem-like and terminally differentiated CD8 T cell subsets previously described in murine chronic viral infection (that is, LCMV)²⁰. Gene set enrichment (GSEA) found that the genes expressed by tumour-infiltrating populations were highly enriched with the analogous cell population described in LCMV (Fig. 2b). We compared these subsets to human effector and memory subsets, and both populations were more similar to the effector cells

than memory²⁹ (Extended Data Fig. 5a–e). These transcriptional data suggest key functional differences between the TCF1⁺ stem-like and TIM3⁺ terminally differentiated T cell subsets within human tumours, and that these functions appear to be similar to what has been described in stem and terminally differentiated CD8 T cells in mice.

To understand how epigenetic mechanisms affect the different functions of these subsets, we performed whole-genome DNA methylation analysis. As T cells underwent transition from naive cells to the stem-like and terminally differentiated states, demethylation events outweighed methylation events approximately 9 to 1 (Fig. 2c, d, Extended Data Fig. 5g–j). These epigenetic changes occurred near key genes involved in differentiation such as *TCF7*, *TBX21*, *PDCD1* and many other checkpoint molecules (Fig. 2e, Extended Data Fig. 5k). Together these data highlight that two key functional characteristics of T cells – proliferative potential and cell killing – are compartmentalized into two distinct populations, and these functions are tightly regulated by transcriptional and epigenetic mechanisms to ensure that cells perform as required.

TCF1⁺ CD8 T cells reside in APC niches

Our finding of a stem-like CD8 T cell population within the tumour, rather than in lymphoid tissue, is unexpected. In mouse models of chronic infection, analogous TCF1⁺ stem-like T cells are found only in lymphoid tissue^{20,21}. Thus, having identified these stem-like cells in tumour tissue, we reasoned that a lymphoid-like microenvironment within the tumour may support their survival in the tumour. We measured tumour-infiltrating antigen-presenting cell (APC) populations (Fig. 3a). This revealed a highly significant correlation – across kidney, prostate and bladder tumours – between the presence of dendritic cells and the number of stem-like CD8 T cells in the tumour (Fig. 3b, Extended Data Fig. 6h). The percentage of macrophages present did not correlate with the presence of TCF1⁺ CD8 T cells or the number of CD8 T cells (Fig. 3b). We then used immunofluorescence staining to determine the spatial relationship between APCs and stem-like CD8 T cells (Fig. 3c, d, Extended Data Fig. 6c–d). TCF1⁺ CD8 T cells were only found in regions with aggregations of major histocompatibility complex II (MHC-II)⁺ cells greater than 5 cells per 10,000 μm^2 (Fig. 3e, f). In contrast, the TCF1[−] population was distributed across the tissue with no preference for APC dense zones (Fig. 3f). We expanded this analysis to large sections of tumour tissue and found that tumours had many regions with dense APC zones, and the stem-like CD8 cells preferentially resided there (Extended Data Fig. 6e–j). When we looked at prostate and bladder tumours, TCF1⁺ CD8 cells were also found in dense APC zones (Fig. 3g, Extended Data Fig. 6k, l). Lastly, we found a significant correlation ($P < 0.05$, $R^2 = 0.73$) between the number of TCF1⁺ CD8 T cells in a tumour and the proportion of the tumour with sufficient APC density to support stem-like cells (Fig. 3g). This suggests that APC dense regions serve as an intra-tumoral niche for stem-like CD8 T cells, which sustain the terminally differentiated T cell population and thus of the anti-tumour immune response.

We next assessed whether these antigen-presenting niches were similar to tertiary lymphoid structures (TLS) previously described in other cancer types^{30,31}. These structures were macroscopically visible with haematoxylin and eosin staining in 5 out of 33 patients, with densely packed mononuclear cells compartmentalized and usually found outside the tumour border (Extended Data Fig. 7a, b). The presence of TLS did not correlate with CD8 T cell infiltration (Extended Data Fig. 7f–h). Visualized using immunofluorescence, TLS were predominantly very densely packed MHC-II⁺ cells, interspersed with few CD8 T cells (Extended Data Fig. 7d). On comparison to human tonsil tissue, these TLS much more closely resembled B cell follicles, which is consistent with several other reports^{30–32} (Extended Data Fig. 7c, d). In comparison, the antigen-presenting niches populated by TCF1⁺ CD8 T cells were predominantly found inside the stromal barrier of the tumour (Extended Data Fig. 7b). Of interest, these nests containing TCF1⁺ CD8

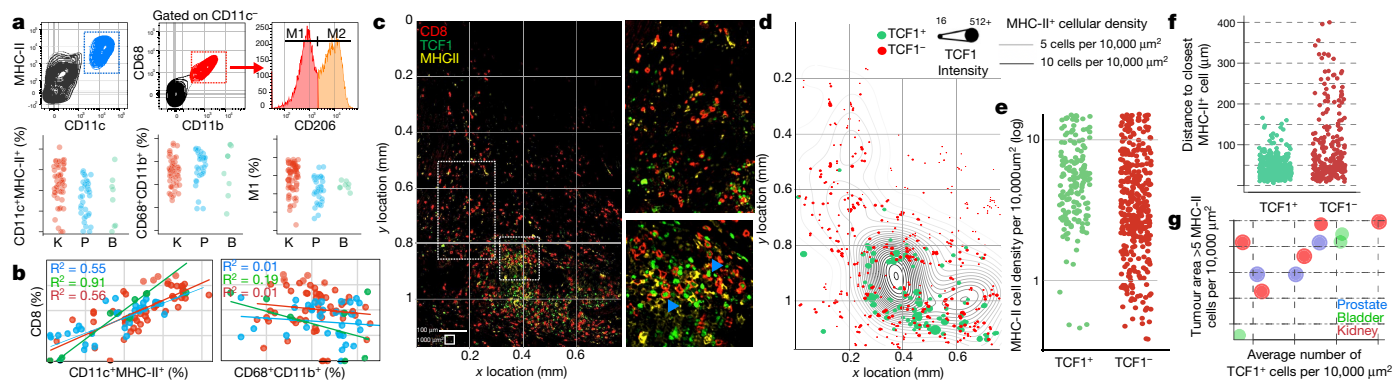


Fig. 3 | APCs form a supportive, intra-tumoral niche for TCF1⁺ stem-like CD8 T cells. **a**, Identification of APC subsets in kidney (red, $n = 53$), bladder (green, $n = 7$) and prostate tumours (blue, $n = 33$). **b**, Correlation between CD8 T cells and APC populations. Percentage of total cells in the tumour that were CD8⁺ T cells and dendritic cells (CD11c⁺MHC-II⁺) or macrophages (CD68⁺CD11b⁺) in patients from **a**. Spearman correlation coefficient is shown. **c**, Immunofluorescence for MHC-II staining identifies APCs, whereas CD8 and TCF1 identify stem-like and terminally differentiated CD8 T cell populations in a representative patient with kidney cancer. Insets show regions highlighted in the larger image. Blue arrows denote examples of TCF1⁺ CD8 T cells. **d**, Cellular spatial relationship map. After acquiring XY coordinates of MHC-II⁺ cells, MHC-

II cellular density was calculated (number of MHC-II⁺ cells per 10,000 μm^2). XY location of CD8 T cells are overlaid with MHC-II density contour. CD8 cells were designated TCF1 positive or negative using histo-cytometry (Extended Data Fig. 6). **e**, MHC-II cellular density surrounding TCF1⁺ or TCF1⁻ subsets. MHC-II density at the corresponding XY coordinates of each CD8 T cell is shown. **f**, Distance between CD8 T cells and the closest MHC-II⁺ cell. **g**, Numerous regions of high MHC-II density correlates within increased number of TCF1⁺ cells in multiple tumour types. y -axis shows proportion of the tumour with MHC-II density >5 MHC-II⁺ cells per 10,000 μm^2 , with average number of TCF1⁺ CD8 T cells in the tumour on the x -axis.

T cells closely resembled the extrafollicular regions of lymphoid tissue where T cells reside – moderately densely arranged APCs packed with many TCF1⁺ CD8 T cells (Extended Data Fig. 7c, e). In addition, we found a significantly higher level of blood and lymphatic endothelial cells (CD31⁺PDPN⁺, CD31⁺PDPN⁺, respectively) in tumours with CD8 infiltration, and these vessels were often closely associated with dense regions of T cell infiltration (Extended Data Fig. 8). Together these findings highlight key features of the CD8 T cell response to cancer. Regions exist in tumours that resemble a T cell zone of lymphatic tissue. These regions contain the TCF1⁺ CD8 T cells that seem to only reside in close proximity to APCs, and the generation of these immune niches is correlated to lymphatic and blood vessel infiltration into the tumour.

Loss of APC niche during immune escape

We next examined how the immune niche differs between patients with controlled disease after surgery compared to those whose tumours escaped immune control and rapidly progressed. We imaged large regions of tumour tissue from 26 patients with kidney cancer at the time of surgery to understand how the presence of immune niches in the tumour might correlate with disease progression (see Extended Data Fig. 9a for patient characteristics). Immunofluorescence quantification of CD8 T cells strongly correlated with flow cytometry quantification of CD8 T cell infiltration (Extended Data Fig. 9b, c). Across around 100,000 20 \times fields of view in these 26 samples, regardless of the level of CD8 T cell infiltration in the patient, we could generally identify a few dense regions of MHC-II where TCF1⁺ CD8 T cells resided (Fig. 4a–d). Most importantly, patients with controlled disease had significantly more of these dense regions (Fig. 4e, f). On stratifying patients above or below the median MHC-II density, we found that patients with low MHC-II⁺ cell density experience significantly impaired progression-free survival (Fig. 4g, $P = 0.04$, HR = 3.157). These factors were independent of PD-L1 expression in the tumour, which had no correlation to the level of CD8 or survival of patients (Extended Data Fig. 10). Importantly, when we specifically studied patients with stage III disease, around 50% of whom progress after surgery, there were >10 -fold fewer immune niches in patients who progressed (Extended Data Fig. 9e–g). Patients with progressive disease also had lower proportions of MHC-II⁺ dense,

CD8⁺ dense, and shared MHC-II⁺ and CD8⁺ dense regions in their tumour (Fig. 4h, i, Extended Data Fig. 9h, i), suggesting that for tumours to evade destruction by CD8 T cells, they must either prevent formation of intra-tumoral immune niches or find ways to destroy them.

Discussion

In this study, we sought to understand the mechanisms controlling CD8 T cell infiltration into human tumours. We found that tumour-infiltrating T cells are comprised of two functionally distinct subsets, a TCF1⁺ stem-like CD8 T cell population, and their progeny, a clonally related terminally differentiated population that express high levels of checkpoint molecules. These terminally differentiated cells fit the traditional definition of an exhausted CD8 T cell; they do not proliferate in response to re-stimulation and express high levels of checkpoint molecules. However, the presence of this terminally differentiated cell population positively correlates with the total number of tumour-infiltrating T cells and protection from disease progression. These observations are not well explained by a model of T cell exhaustion whereby continuous antigen exposure leads to accumulation of checkpoint molecules, resulting in a decline of the T cell response. Based on the functional characteristics we defined in these two cell populations and on the clonal relationship between stem-like and terminally differentiated cells, we propose that the stem-like CD8 T cell acts as a precursor to generate a terminally differentiated effector population, which is in agreement with other previous studies^{22–27}. In this model, the stem-like cells require a region within the tumour that resembles the T cell zone of secondary lymphatic tissues, made up of dense areas of antigen-presenting cells. An unanswered question in this model is how stem-like CD8 T cells originate in the tumour. Previous studies have shown that CD8 T cells in tissue-draining lymph nodes are transcriptionally and phenotypically similar to the stem-like CD8 T cell described in chronic LCMV infection, suggesting this may be the source of the stem-like cells in tumours³³.

On the basis of this model, we propose that the decline of the T cell response in human cancer is not caused by accumulation of checkpoint-expressing exhausted CD8 T cells or overexpression of PD-L1 in the tumour, but by the failure of stem-like CD8 T cells to be sufficiently

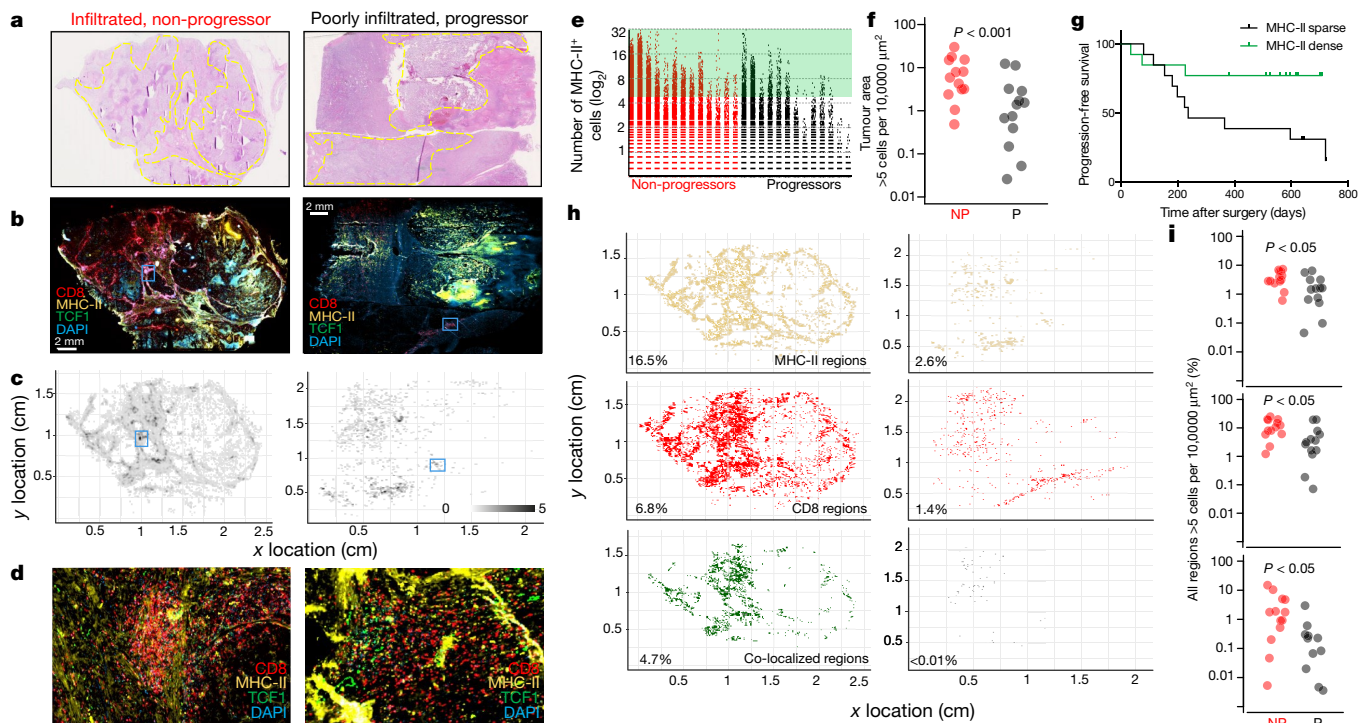


Fig. 4 | Loss of APC niche is associated with impaired CD8 T cell response and disease progression. **a–d**, Patients with dense T cell infiltration and no disease progression (red, left) and one with poor T cell infiltration and progressive disease (grey, right). **a**, Haematoxylin and eosin whole-slide images. Tumour is outlined in yellow. **b**, Whole-slide immunofluorescence images. MHC-II (yellow), TCF1 (green), CD8 (red) and DAPI (blue). **c**, Immunomap of APC density in tumours from **b** constructed as in Fig. 3 and Extended Data Fig. 7. **d**, Insets show highlighted regions from **b** and **c**, illustrating regions of high MHC-II density and stem-like T cell infiltration in kidney tumours. **e**, Comparison of the number of MHC-II⁺ cells per 300 $\mu\text{m} \times 300 \mu\text{m}$ field in patients with ($n = 13$)

and without ($n = 13$) progressive disease. **f**, Comparison of the proportion of tumour area with >5 MHC-II⁺ cells per $10,000 \mu\text{m}^2$ between patients with and without progressive disease. Mann–Whitney test result is shown. **g**, Patients with high MHC-II⁺ cell density had improved progression free survival. log-rank statistical analysis yields $P = 0.04$ and $\text{HR} = 3.226$. **h**, Immunomaps illustrating regions of MHC-II⁺ cell density (yellow), CD8⁺ cell density (red) or shared density (green) in tumours from **a–d**. **i**, Patients without progressive disease have more areas where the density of MHC-II⁺ cells (top), CD8⁺ cells (middle) or MHC-II⁺ and CD8⁺ cells (bottom) exceeds 5 cells per $10,000 \mu\text{m}^2$.

stimulated by an antigen-presenting-cell niche to continuously produce terminally differentiated CD8 T cells in the tumour. Furthermore, the scarcity of these niches in tumours that rapidly progress after surgery suggests that tumours may be interfering with the formation or continued maintenance of immune niches and that this may be a novel mechanism of immune evasion requiring further investigation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1836-5>.

- Galon, J. et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* **313**, 1960–1964 (2006).
- Pagès, F. et al. Effector memory T cells, early metastasis, and survival in colorectal cancer. *N. Engl. J. Med.* **353**, 2654–2666 (2005).
- Peranzoni, E. et al. Macrophages impede CD8 T cells from reaching tumor cells and limit the efficacy of anti-PD-1 treatment. *Proc. Natl Acad. Sci. USA* **115**, E4041–E4050 (2018).
- Azimi, F. et al. Tumor-infiltrating lymphocyte grade is an independent predictor of sentinel lymph node status and survival in patients with cutaneous melanoma. *J. Clin. Oncol.* **30**, 2678–2683 (2012).
- Savas, P. et al. Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. *Nat. Med.* **24**, 986–993 (2018).
- Herbst, R. S. et al. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature* **515**, 563–567 (2014).
- Tumeh, P. C. et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* **515**, 568–571 (2014).
- Eroglu, Z. et al. High response rate to PD-1 blockade in desmoplastic melanomas. *Nature* **553**, 347–350 (2018).

- Gallimore, A., Dumrese, T., Hengartner, H., Zinkernagel, R. M., Rammensee, H.-G. Protective immunity does not correlate with the hierarchy of virus-specific cytotoxic T cell responses to naturally processed peptides. *J. Exp. Med.* **187**, 1647–1657 (1998).
- Zajac, A. J. et al. Viral immune evasion due to persistence of activated T cells without effector function. *J. Exp. Med.* **188**, 2205–2213 (1998).
- Wherry, E. J. et al. Molecular signature of CD8⁺ T cell exhaustion during chronic viral infection. *Immunity* **27**, 670–684 (2007).
- Barber, D. L. et al. Restoring function in exhausted CD8 T cells during chronic viral infection. *Nature* **439**, 682–687 (2006).
- Gros, A. et al. PD-1 identifies the patient-specific CD8⁺ tumor-reactive repertoire infiltrating human tumors. *J. Clin. Invest.* **124**, 2246–2259 (2014).
- Topalian, S. L. et al. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N. Engl. J. Med.* **366**, 2443–2454 (2012).
- Brahmer, J. R. et al. Safety and activity of anti-PD-L1 antibody in patients with advanced cancer. *N. Engl. J. Med.* **366**, 2455–2465 (2012).
- Hodi, F. S. et al. Improved survival with ipilimumab in patients with metastatic melanoma. *N. Engl. J. Med.* **363**, 711–723 (2010).
- Ahmadzadeh, M. et al. Tumor antigen-specific CD8 T cells infiltrating the tumor express high levels of PD-1 and are functionally impaired. *Blood* **114**, 1537–1544 (2009).
- Mlecik, B. et al. Integrative analyses of colorectal cancer show immunoscore is a stronger predictor of patient survival than microsatellite instability. *Immunity* **44**, 698–711 (2016).
- Tosolini, M. et al. Clinical impact of different classes of infiltrating T cytotoxic and helper cells (T_H1 , T_H2 , T_{reg} , T_H17) in patients with colorectal cancer. *Cancer Res.* **71**, 1263–1271 (2011).
- Im, S. J. et al. Defining CD8⁺ T cells that provide the proliferative burst after PD-1 therapy. *Nature* **537**, 417–421 (2016).
- He, R. et al. Follicular CXCR5-expressing CD8⁺ T cells curtail chronic viral infection. *Nature* **537**, 412–416 (2016).
- Utzschneider, D. T. et al. T cell factor 1-expressing memory-like CD8⁺ T cells sustain the immune response to chronic viral infections. *Immunity* **45**, 415–427 (2016).
- Brummelman, J. et al. High-dimensional single cell analysis identifies stem-like cytotoxic CD8⁺ T cells infiltrating human tumors. *J. Exp. Med.* **215**, 2520–2535 (2018).
- Sade-Feldman, M. et al. Defining T cell states associated with response to checkpoint immunotherapy in melanoma. *Cell* **175**, 998–1013 (2018).
- Siddiqui, I. et al. Intratumoral Tcf1⁺PD-1⁺CD8⁺ T cells with stem-like properties promote tumor control in response to vaccination and checkpoint blockade immunotherapy. *Immunity* **50**, 195–211 (2019).

26. Kurtulus, S. et al. Checkpoint blockade immunotherapy induces dynamic changes in PD-1⁺CD8⁺ tumor-infiltrating T cells. *Immunity* **50**, 181–194 (2019).
27. Miller, B. C. et al. Subsets of exhausted CD8⁺ T cells differentially mediate tumor control and respond to checkpoint blockade. *Nat. Immunol.* **20**, 326–336 (2019).
28. Simoni, Y. et al. Bystander CD8⁺ T cells are abundant and phenotypically distinct in human tumour infiltrates. *Nature* **557**, 575–579 (2018).
29. Akondy, R. S. et al. Origin and differentiation of human memory CD8 T cells after vaccination. *Nature* **552**, 362–367 (2017).
30. Dieu-Nosjean, M. C., Goc, J., Giraldo, N. A., Sautès-Fridman, C. & Fridman, W. H. Tertiary lymphoid structures in cancer and beyond. *Trends Immunol.* **35**, 571–580 (2014).
31. Sautès-Fridman, C., Petitprez, F., Calderaro, J. & Fridman, W. H. Tertiary lymphoid structures in the era of cancer immunotherapy. *Nat. Rev. Cancer* **19**, 307–325 (2019).
32. Silina, K. et al. Germinal centers determine the prognostic relevance of tertiary lymphoid structures and are impaired by corticosteroids in lung squamous cell carcinoma. *Cancer Res.* **78**, 1308–1320 (2017).
33. Miron, M. et al. Human lymph nodes maintain TCF-1^{hi} memory T cells with high functional potential and clonal diversity throughout life. *J. Immunol.* **201**, 2132–2140 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

Sample collection, preparation and storage

Patients were recruited in accordance with an approved IRB protocol, and all patients provided informed consent. Patient tumour samples were collected immediately after undergoing partial or radical nephrectomy or prostatectomy or undergoing transurethral resection of a bladder tumour (TURBT). Samples for flow cytometric analysis were harvested in Hank's Balanced Salt Solution, minced into small pieces, digested using Liberase enzyme cocktail, and homogenized using a MACS Dissociator. Single cell suspensions were obtained, RBC ACK lysed, and stored at -80°C in freezing media for batch analysis. Samples for immunofluorescence analysis were formaldehyde fixed and embedded in paraffin blocks by Emory Pathology. Unstained and haematoxylin/eosin stained sections of FFPE blocks were obtained from Emory Pathology.

Statistical analysis

Patients were selected to have at minimum 365 days of follow up. Follow up time was calculated as the number of days from the date of surgery to an event or to censorship. Progression and death were classified as events. Patients who had not progressed or are not deceased were censored, and the number of days is calculated from the date of surgery to 9 May 2018. Investigators were not blinded during outcome assessment. Statistical analysis was conducted using GraphPad Prism or using SAS Version 9.4 and SAS macros developed by the Biostatistics and Bioinformatics Shared Resource at Winship Cancer Institute. The significance level was set at $P < 0.05$. Descriptive statistics for each variable were reported. The univariate association with percentage of CD8 T cells was carried out by ANOVA/Kruskal–Wallis test for categorical covariates and by Pearson correlation coefficient for numerical covariates. The univariate association of each covariate with PFS was tested by proportional hazard model with hazard ratio and its 95% confidence interval being reported. We examined a possible nonlinear relationship between a continuous percentage of CD8 T cells and PFS through a martingale residual plot and identified an optimal cut-off value of percentage of CD8 T cells that maximizes the separation between the two groups by a bias adjusted log rank test^{34,35}. The method enables the estimation and evaluation of the significance of the cut-off value and also is adjusted for the bias created by the data driven searching process. The optimal cut-off value was found to be 2.2% (Extended Data Fig. 1b & 1c). Using this same 2.2% cut-off for CD8 infiltration in patients with more aggressive, non-metastatic disease (T3N0M0), less CD8 T cell infiltration predicted a sixfold more rapid progression (Extended Data Fig. 1d). CD8 T cell infiltration also significantly predicted progression amongst patients categorized as high-risk by a conventional prognostic scoring system (SSIGN) (Extended Data Fig. 1e). Statistical methods were not used to pre-determine number of patients included.

Flow cytometry

Single cell suspensions from human tumours were stained with antibodies listed in Supplementary Information Table 1. Live/dead discrimination was performed using fixable Near-IR Dead Cell Stain Kit (Invitrogen). Samples were acquired with a Becton Dickinson LSR II and analysed using FlowJo. For intracellular staining, cells were fixed and permeabilized using the FOXP3 Transcription Factor Staining Buffer Set (eBioscience).

Proliferation assays

CD8 T cells subsets were sorted from tumours and labelled with Cell trace violet (Thermo) according to manufacturer's instruction. Cells were incubated with anti-CD3/anti-CD28 T cell activation beads (Miltenyi) at a ratio of 1 bead to 2 T cells in U-bottom plates. 10 U ml^{-1} of human IL-2 (Peprotech) was included in culture media (RPMI + 10% FBS). After 4 days, cells were analysed by flow cytometry for proliferation and expression of various proteins. Proliferation index was assessed using FlowJo.

In vitro assays

Stem-like and terminally differentiated CD8 T cells were sorted from human tumours and incubated with T cell culture media (RPMI + 10%FBS) supplemented with human IL-2 (10 IU ml^{-1}) in U-bottom plates. After 3 days, cells were analysed by flow cytometry for expression of various proteins.

TCR sequencing

Single cell suspensions from human tumours were stained with antibodies listed in Supplementary Information Table 1. Live/dead discrimination was performed using fixable Near-IR Dead Cell Stain Kit (Invitrogen). Populations of interest were isolated using a Becton Dickinson FACS Aria II Cell Sorter. Gating is shown in Extended Data Fig. 4a, c. DNA was isolated using a Qiagen AllPrep DNA/RNA Micro Isolation Kit. TCR sequencing was performed by Adaptive Biotechnologies Immunoseq technologies. TCR Sequencing analysis was performed using custom R scripts. The number of TCRs detected and degree of overlap detected was highly subject to the number of cells collected, highlighting the need to sufficiently sample the pools of cells to accurately understand the clonal relationship between them (Extended Data Fig. 4e, f).

To determine if there was significant overlap between populations, we first calculated the contamination of each population with the other so we could determine if overlap in TCRs could be explained by the contamination rate. To determine the overlap between the stem and terminally differentiated cells due to biological and technical variance, flow cytometry data was fit using an EM mixing model³⁶. The characteristics of these fitted models are shown in Extended Data Fig. 4b. Shown on the plot are 80% and 95% confidence intervals for each population and the approximate position of gates used to sort populations. We then placed gates where we had for the sort and asked the question of what proportion of the cells in that gate were derived from the target and contaminating population. This contamination rate is highly subject to the ratio of the two populations. In our 2 most extreme patients shown in Extended Data Fig 4e, if 93% of the cells are the stem-like population, the contamination rate in the TD population is as high as 14%.

Extended Data Figure 4b shows how the purity changes as the ratio of stem to terminally differentiated cells changes. The two most extreme samples are highlighted on the figure to show what the inferred proportion of each population is in the sorted cells. In addition, we added 5% to this number for each sample to account for additional contamination from the sorting procedure. The summary of this analysis is included in Extended Data Fig. 4h.

To identify significance of TCR overlap we used the purity calculated for each patient we tested if the relative frequency of each TCR could be explained by contamination. For each specific TCR that was detected in both populations, we tested two hypotheses. First, can the number of a particular TCR in the stem-like population be accounted for by contamination from the TD cell population, and conversely, can the same TCR in the TD population be accounted for by contamination from the stem-like population. This was achieved by assuming each TCR detected in a sample was a Bernoulli trial with a probability of occurring equal to the expected frequency of the TCR due to contamination. For example, we assumed that if a TCR was found at a frequency of 10% in the stem population, and the inferred overlap into the TD was 10%, it would contaminate the terminally differentiated cells at a frequency of 1%. If we collected 1000 total TCRs for a particular sample, and detected 10 of this specific TCR, the probability of detecting at least this many TCRs due to this 1% contamination rate would be given by:

$$P(X \leq 10) = \sum_{i=0}^{10} \binom{1000}{i} 0.01^i (1 - 0.01)^{1000-i}$$

The general formula for testing if the overlap in the terminally differentiated population is caused by contamination from the stem-like cells is given by:

$$P(X \leq k) = \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{(n-i)}$$

Where k = number of the specific TCR detected in the terminally differentiated population; P = frequency of the specific TCR in the stem population x contamination rate; n = total number of TCRs detected in a sample.

We applied this analysis to every TCR collected that had overlap detected and tested the converse hypothesis that the fraction of stem-like TCRs detected could be accounted for by contamination from the terminally differentiated cells. If both tests were under 0.05, we rejected the hypothesis that the overlap was caused by contamination. Figure 1l highlights the proportion of TCRs in each sample that meet these criteria. The supplementary table (Extended Data Fig. 4h) provided these values used for every TCR and the P value calculated.

To identify significance of TCR overlap, we assumed 90% purity and conducted a Fisher's exact test to test the hypothesis that the TCR overlap we detected could be explained by this contamination rate. To determine the probability that an overlap could have been detected given the number of cells recovered, we fit an exponential distribution of the observed stem and effector TCR clone frequency (shown in Extended Data Fig. 4b). We then used a bootstrapping approach to randomly sample the same number of TCRs from these two distributions as cells we had collected. We repeated this 1,000 times. If a 20% overlap was not detected at least 80% of the time, the sample was considered underpowered to detect an overlap. Analysis of the TCRs found that the TCR repertoires showed a high degree of immunodominance, where the ten most dominant clones account for 55% of the terminally differentiated repertoire and for 31% of the stem-like repertoire, indicating an expansion against a narrow range of antigens in the tumours (Extended Data Fig. 4d).

RNA sequencing and analysis

RNA was isolated from FACS sorted cells using QIAGEN All-prep kit. RNA was prepared using Contech SmartSeq2 (Bladder samples) or Nugen Ovation (Prostate, Kidney samples) library prep kits. Prostate and Kidney samples were sequenced at HudsonAlpha on a HiSeq25000, Bladder samples were sequenced at the Emory Yerkes Genomics Core on a HiSeq1000. Data was normalized and differential expression of genes identified using DESeq2³⁷. Raw fastq files and analysis of RNA-seq is uploaded to GEO under identifier GSE140430.

DNA-methylation analysis

Whole-genome DNA methylation was performed using the Illumina TruSeq DNA Methylation Kit. Sequence data was aligned using Bismark³⁸, and data was analysed using custom R and Python scripts which are available upon request. Briefly, individual significantly differentially methylated CpG motifs were identified by Fisher's exact test. Continuous regions of differentially methylated CpGs were identified by finding regions where at least 6 out of 10 CpGs in a continuous stretch were differentially methylated. These regions were then collapsed and analysed as single 'differentially methylated regions' (DMRs). Differentially expressed regions were identified as those that had a p value less than 1×10^{-4} by Fisher's exact test and were at least 20% different to the comparison sample. Transcription factor binding enrichment analysis was also conducted, identifying TCF4, TCF7L2, and MYC as enriched in the stem-like cells and E2F, NRF2, and SP1 in the terminally differentiated cells (Extended Data Fig. 5l). Whole-genome DNA methylation data are uploaded to GEO under identifier GSE140430.

Deparaffinization and antigen retrieval

Sections were deparaffinized in successive incubations with xylene and decreasing concentrations (100, 95, 75, 50, 0%) of ethanol. Antigen retrieval was achieved using either Abcam 100× Citrate Antigen

Retrieval Buffer (pH = 6.00) for 20 min at 100 °C, followed by 20 min at ambient temperature or Abcam 100× TrisEDTA Antigen Retrieval Buffer (pH = 9) heated to 115 °C under high pressure. Sections were then washed in either a solution of 10 mM glycine and 0.2% sodium azide in phosphate buffered saline or PBS + 0.1% Tween20 before antibody staining.

Immunofluorescence antibody staining

Sections were blocked for 15–30 min with a 5% goat serum, 1% bovine serum albumin blocking solution containing 10 mM glycine and 0.2% sodium azide or PBS + 0.1% Tween20. Sections were then stained with appropriate primary and secondary antibodies. Primary antibodies were used at a concentration of 1:100 and incubated for 1 h at room temperature. Secondary antibodies were used at a concentration of 1:250 and incubated for 30 min at room temperature. Detailed information about antibodies used is listed in Supplementary Information Table 2.

PD-L1 staining and scoring

FFPE slides for 45 patients were stained using Agilent Biotechnologies PD-L1 IHC (clone 22C3 pharmDx) Staining Kit by Emory Pathology Laboratories. Clinical-grade scoring of PD-L1 status was performed by two board-certified pathologists at Emory University Hospital. Slides with 1–49% of tumour cells expressing PD-L1 were scored 'positive-low', slides with 50+% of tumour cells expressing PD-L1 were scored 'positive-high', and slides with <1% of tumour cells expressing PD-L1 were scored 'negative.'

Image capture and analysis

We selected a fluorophore panel which allowed for simultaneous visualization of three targets and a nuclear stain (DAPI). For images shown in Fig. 3, we used a Leica SP8 confocal microscope with a motorized stage for tiled imaging, and a 40x, 1.3NA, 0.24 mm WD oil immersion objective was used, allowing for highly resolved, smoothly tiled images. Fluorophores were excited with the 496, 561, and 594 laser lines or with a multiphoton Coherent Chameleon Vision II laser, tuned to 700 nm (DAPI). Emission-optimized wavelength ranges informed specific detector channels, which were used to detect fluorescence. Leica LASX software was used to create a maximum projection image, allowing us to obtain large tiled images regardless of a varying focal plane across each tissue section. For images shown in Fig. 4, we used a Zeiss Z.1 Slide Scanner equipped with a Colibri 7 Flexible Light Source. Zeiss ZenBlue software was used for post-acquisition image processing. For brightfield imaging, slides were scanned using a Hamamatsu's Nanozoomer slide scanner.

CellProfiler, a free, open-source software for image analysis, was used for subsequent image manipulations. CellProfiler was used to define 'primary objects' within images, based upon user-defined parameters (diameter, fluorescence intensity, object clumping, etc.). We used this technique to define DAPI 'primary objects' (that is, all cells) and MHC+ 'primary objects' (that is, defining antigen presenting cells). We also used this technique to define CD8+ 'primary objects,' which we then used to create 'secondary objects' by extending the border of each object by 1 pixel in all directions. These CD8+ 'secondary objects' were used to define CD8+ T cells. Detailed review of parameters used to MHC-II+ antigen presenting cells and CD8+ T cells can be found in Supplementary Information Table 3. We then used CellProfiler to measure the intensity of TCF1 staining intensity in each CD8+ T cell object. Data exported from the CellProfiler pipeline included XY location of CD8+ objects, MHC-II+ objects, and mean intensity of TCF1 staining in CD8+ T cell objects. The remainder of image analysis was carried out using custom R and python scripts. MHC-II density and distance to nearest MHC-II+ neighbour were calculated in custom python scripts.

In order to determine the area of tissue necessary to be sampled to obtain an accurate and quantitative assessment of the CD8 T cell infiltration into tumours, large slide scanned images were dissected into

areas the approximate size of a 20× field of view. Increasing number of random fields of view were sampled from images and the percent of cells that were CD8 positive by immunofluorescence correlated to FACS from the corresponding sample. The estimated number of 20× fields of view necessary to obtain an accurate assessment of level of CD8 T cell infiltration is 171 fields of view (Extended Data Fig. 9d). Histo-cytometric analysis approach employed similar to that reported previously³⁹.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Raw fastq files and associated RNA and whole genome bisulphite sequencing have been uploaded to the NCBI Gene Expression Omnibus (GEO) database under identifier GSE140430. Other relevant data are available from the corresponding author upon reasonable request.

Code availability

Custom code for RNA-seq, whole genome methylation, and quantitative immunofluorescence are available from the corresponding author upon reasonable request.

34. Mandrekar J., Cha, S. Cutpoint determination methods in survival analysis using SAS2003. SAS <https://support.sas.com/resources/papers/proceedings/proceedings/sugi28/261-28.pdf> (2003).
35. Contal, C. & O'Quigley, J. An application of changepoint methods in studying the effect of age on survival in breast cancer. *Comput. Stat. Data Anal.* **30**, 253–270 (1999).
36. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *R J.* **8**, 289–317 (2016).
37. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
38. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics* **27**, 1571–1572 (2011).

39. Gerner, M. Y., Kastenmuller, W., Ifrim, I., Kabat, J. & Germain, R. N. Histo-cytometry: a method for highly multiplex quantitative tissue imaging analysis applied to dendritic cell subset microanatomy in lymph nodes. *Immunity* **37**, 364–376 (2012).

Acknowledgements This work was supported by funding from the Prostate Cancer Foundation, Swim Across America, the James M. Cox Foundation and James C. Kennedy, pilot funding from the Winship Cancer Institute supported by the Dunwoody Country Club Senior Men's Association, and NCI grants 1-R00-CA197891 (H.K.) and U01-CA113913 (M.G.S.). We recognize Adaptive Biotechnologies for providing laboratory services as a part of an educational grant award. We would like to acknowledge the Yerkes NHP Genomics Core which is supported in part by NIH P51 OD011132, the Emory Flow Cytometry Core supported by the National Center for Georgia Clinical & Translational Science Alliance of the National Institutes of Health under award number UL1TR002378, the Intramural Research Program of the NIH, National Cancer Institute and the Emory University Integrated Cellular Imaging Microscopy Core of the Winship Cancer Institute of Emory University and NIH/NCI under award number 2P30CA138292-04.

Author contributions C.S.J. and H.K. conceived and designed the study and composed the manuscript. C.S.J., V.A.M., M.G.S., C.P.F., M.A. and H.K. designed experiments. C.S.J., N.P., J.W.C., M.C., R.M.V., W.H.H., D.M., K.M., A.I.K., K.K., Y.M.C., A.K., A.O.K., A.R. and H.K. collected flow cytometry data. C.S.J., N.P., M.C. and H.K. analysed flow cytometry data. C.S.J., N.P., M.C., A.R. and H.K. performed fluorescence activated cell sorting. C.S.J., N.P. and M.C. performed RNA and DNA extractions. C.S.J., S.W., R.L. and A.G.S. optimized and performed immunofluorescence slide scanning. C.S.J. and H.K. developed quantitative immunofluorescence techniques and performed quantitative analysis of immunofluorescence data. W.H.H., D.M. and H.K. performed RNA sequencing analysis. N.P., M.C., K.K., Y.M.C., A.K. and H.K. performed in vitro T cell assays. N.P. and H.K. performed whole-genome methylation analysis. R.A., S.J.I. and A.O.K. provided critical expertise and contributed specific analysis. V.A.M., M.G.S., C.P.F. and M.A. provided clinical samples. C.S.J., J.W.C. and A.R. collected and organized clinical data. K.M., A.O.O., P.M. and C.E. provided annotation and scoring of pathology specimen. Y.L. assisted with biostatistical analysis. All authors reviewed the manuscript.

Competing interests The authors declare no competing interests.

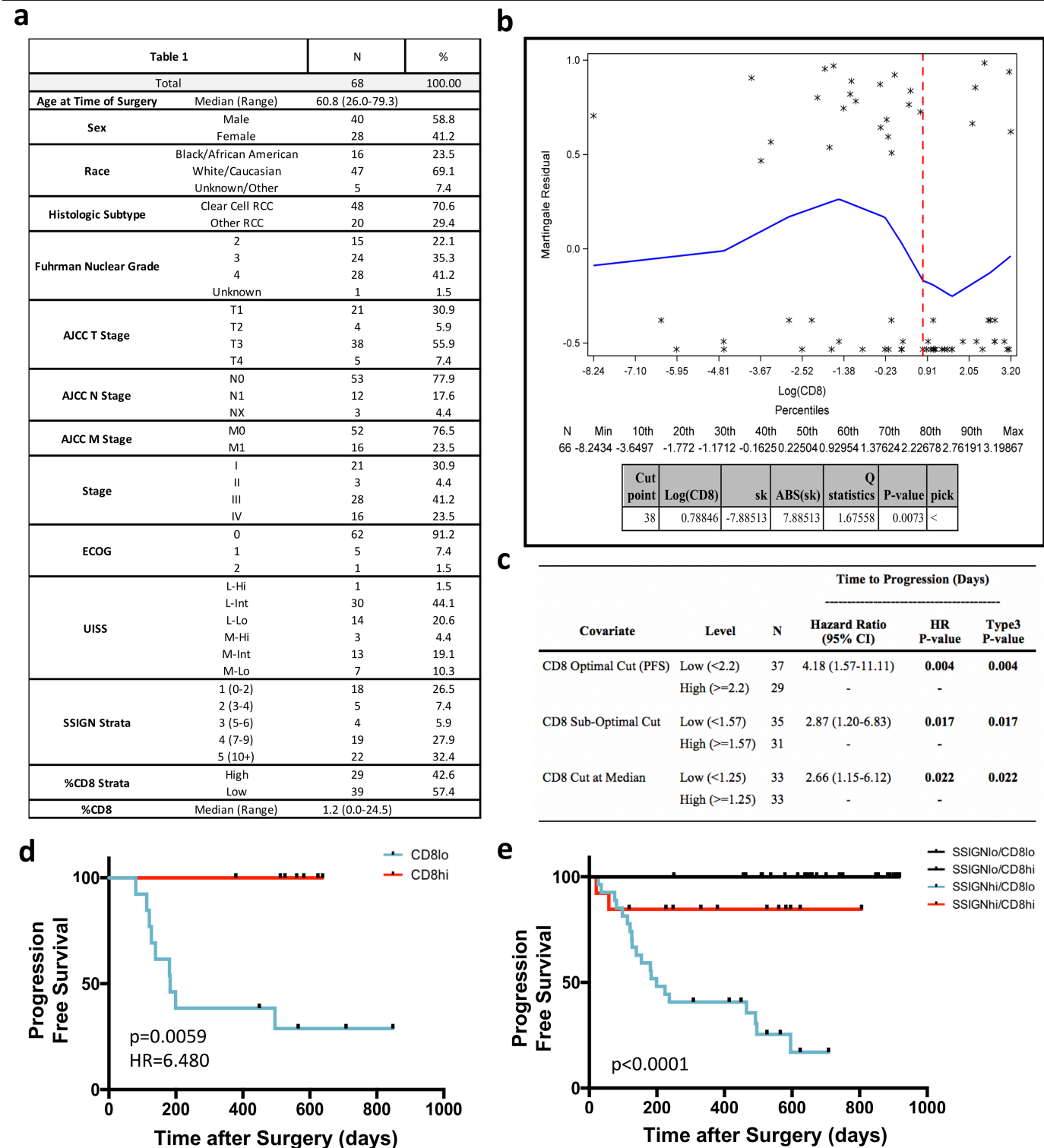
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1836-5>.

Correspondence and requests for materials should be addressed to H.K.

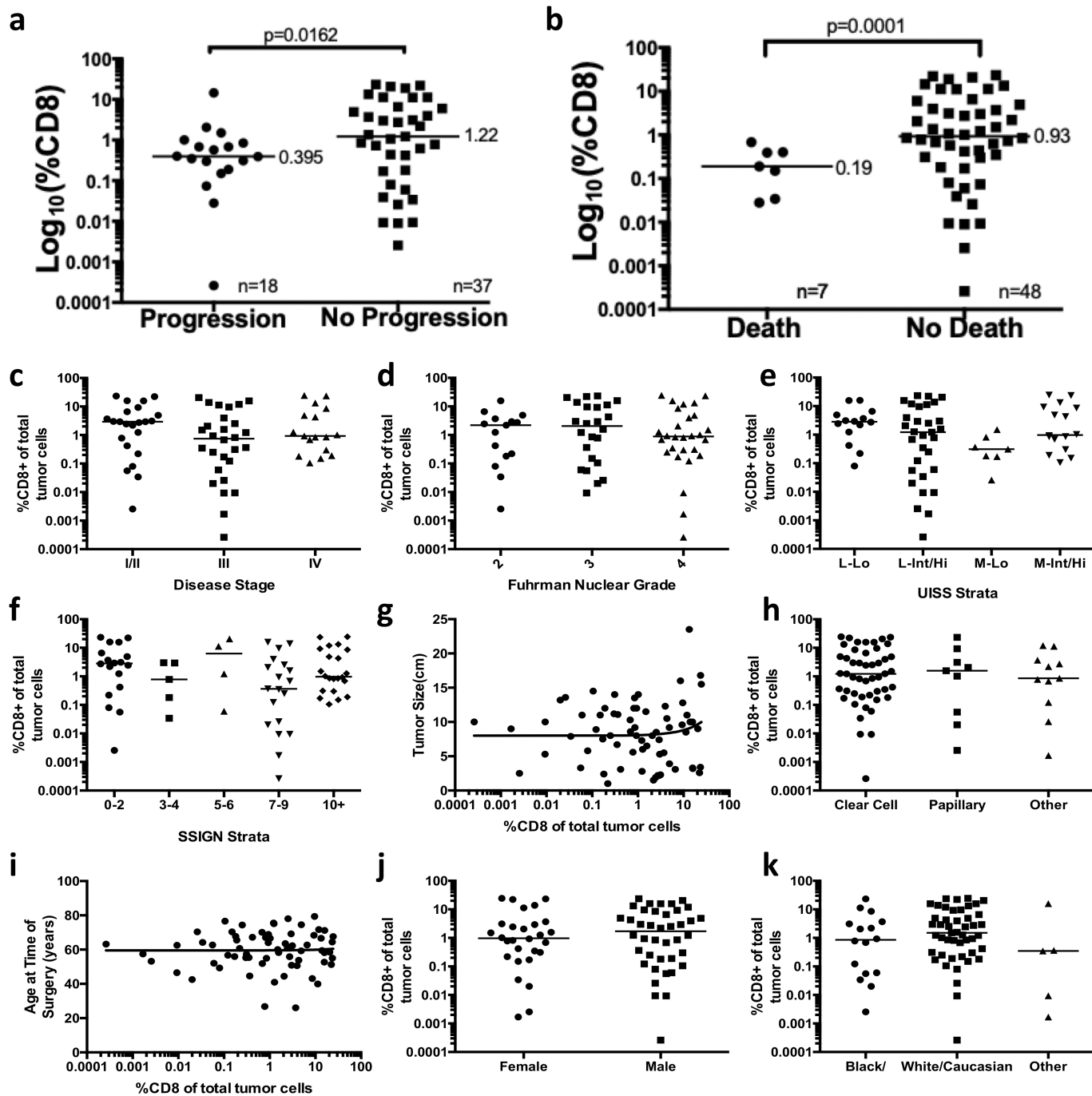
Peer review information Nature thanks I. Mellman, N. P. Restifo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



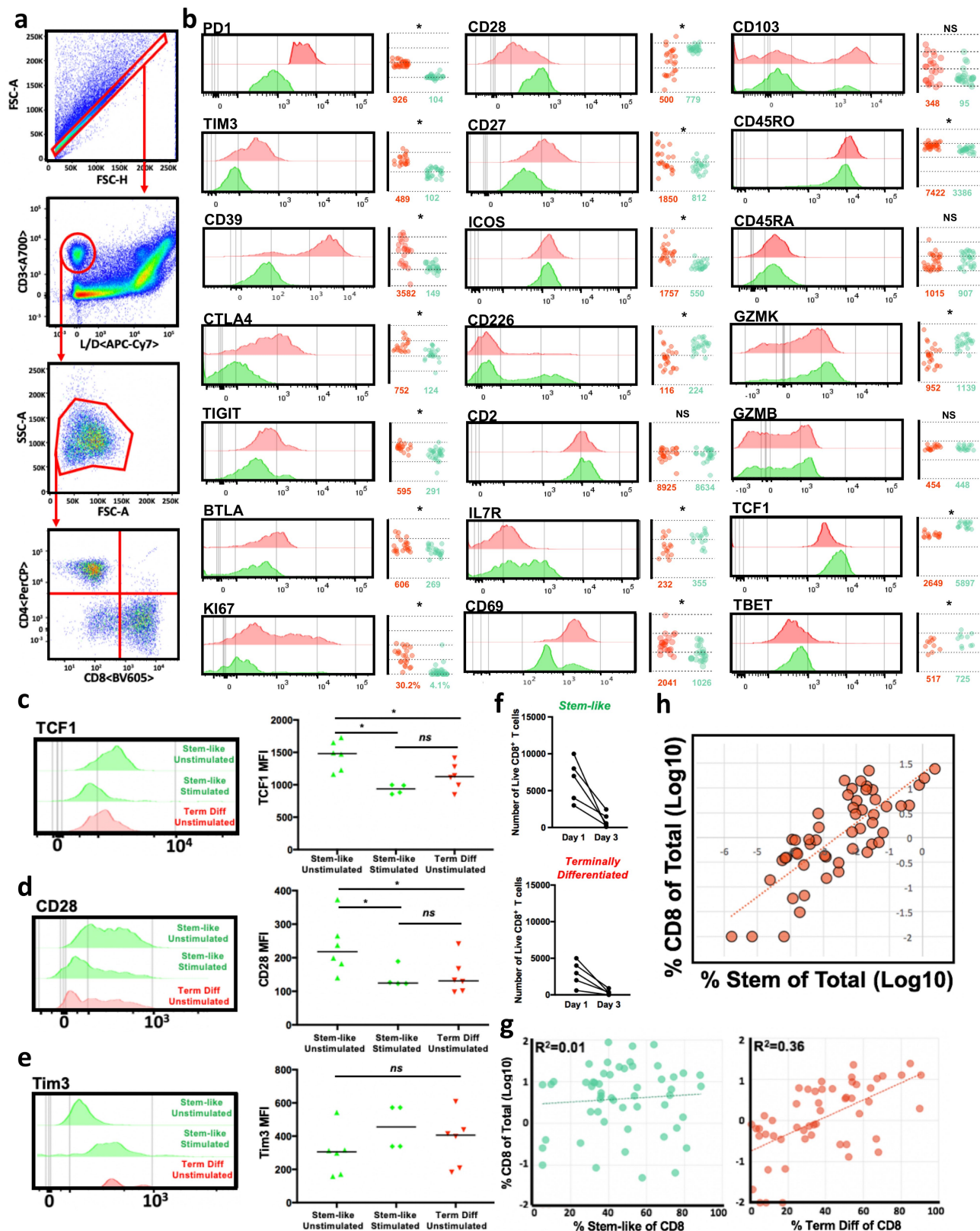
Extended Data Fig. 1 | Description of statistics and sub-group analyses of progression-free survival. **a**, Descriptive statistics. Table details the demographic, disease stage, disease characteristic and immune infiltrate breakdown of the cohort of patients with kidney cancer. **b**, Martingale residual plot illustrating discovery of 2.2% CD8 'optimal cut'. **c**, Comparison of optimal cut, sub-optimal cut and median cut. **d**, CD8 T cell infiltration predicts time to progression in stage III (T3N0M0) patients. Patients were stratified into high (>2.2% CD8) or low (<2.2% CD8) based on the optimal cut identified in a cohort

of all-stage patients. CD8^{hi}, $n=13$; CD8^{lo}, $n=7$. $P=0.0059$, $HR=6.480$, as determined using log-rank test. **e**, CD8 T cell infiltration significantly improves prognostication in patients with kidney cancer with high SSIGN (size, stage, grade, necrosis) scores. $P\leq 0.0001$, as determined using log-rank test. Patients were stratified into low (scores 1–6) and high (scores >6) SSIGN score groups and into low (<2.2% CD8) and high (>2.2% CD8) T cell infiltration. SSIGN^{lo}CD8^{lo}, $n=11$, SSIGN^{lo}CD8^{hi}, $n=16$, SSIGN^{hi}CD8^{lo}, $n=28$, SSIGN^{hi}CD8^{hi}, $n=13$.



Extended Data Fig. 2 | CD8 T cell infiltration is associated with improved survival and is independent of standard risk assessment tools, tumour features and patient demographics. a, b, Proportion of CD8 T cells in the tumours of patients that progress or die after surgery as compared to those without disease progression (a) or death (b). **c,** Disease stage, $P=0.6$.

d, Fuhrman nuclear grade, $P=0.4$. **e,** UISS groups, $P=0.3$. **f,** SSIGN groups, $P=0.3$. **g,** Maximum tumour size in one dimension, in centimetres, $R=0.01$, $P=0.3$. **h,** Histologic subtype, $P=0.7$. **i,** Patient age at the time of surgery, in years, $R=0.001$, $P=0.9$. **j,** Patient sex, $P=0.8$. **k,** Patient race/ethnicity, $P=0.7$. Median value is shown for **a-f, h and j-k**.

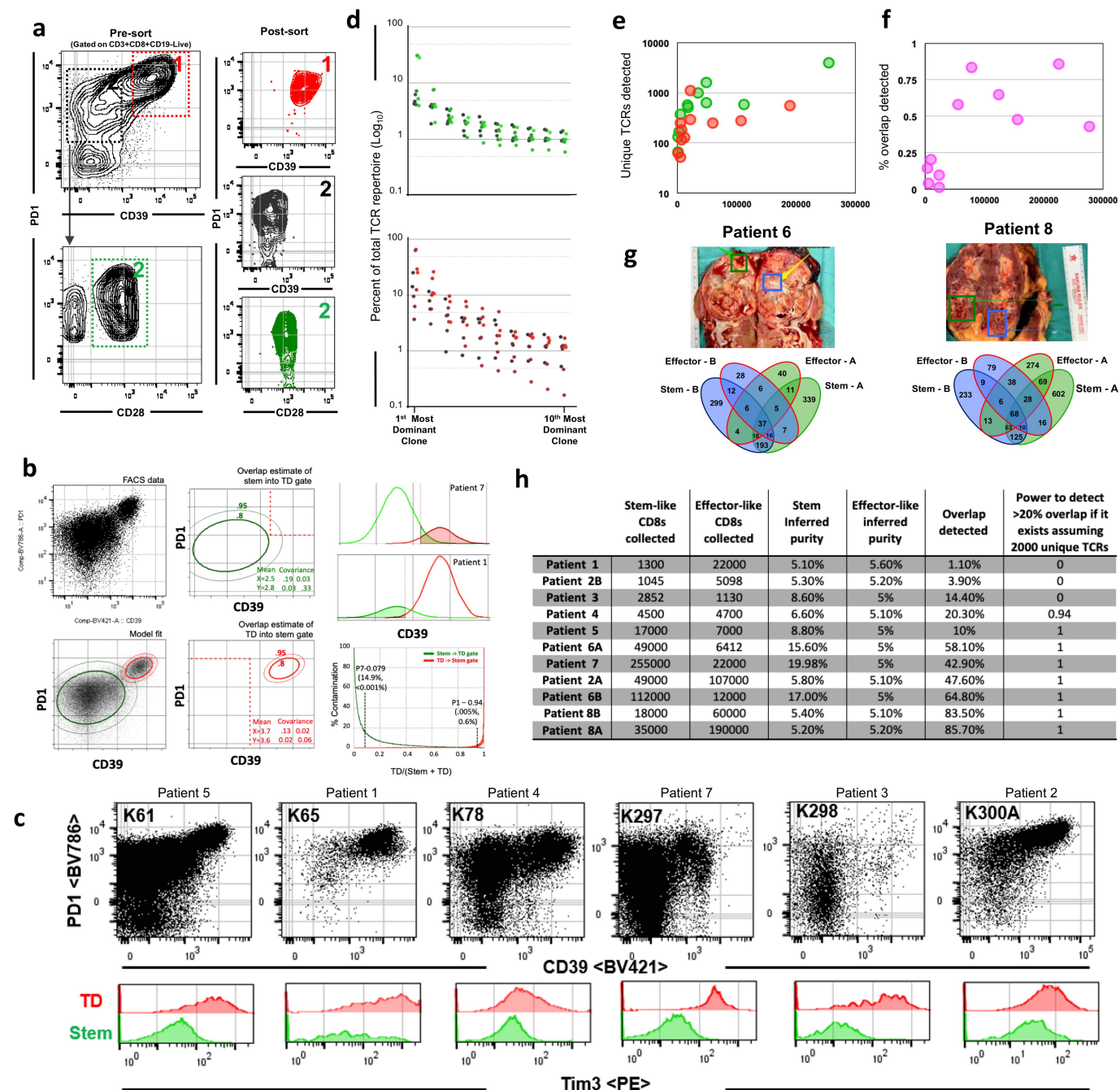


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Flow cytometric comparison and in vitro functional studies of stem-like and terminally differentiated CD8 T cells.

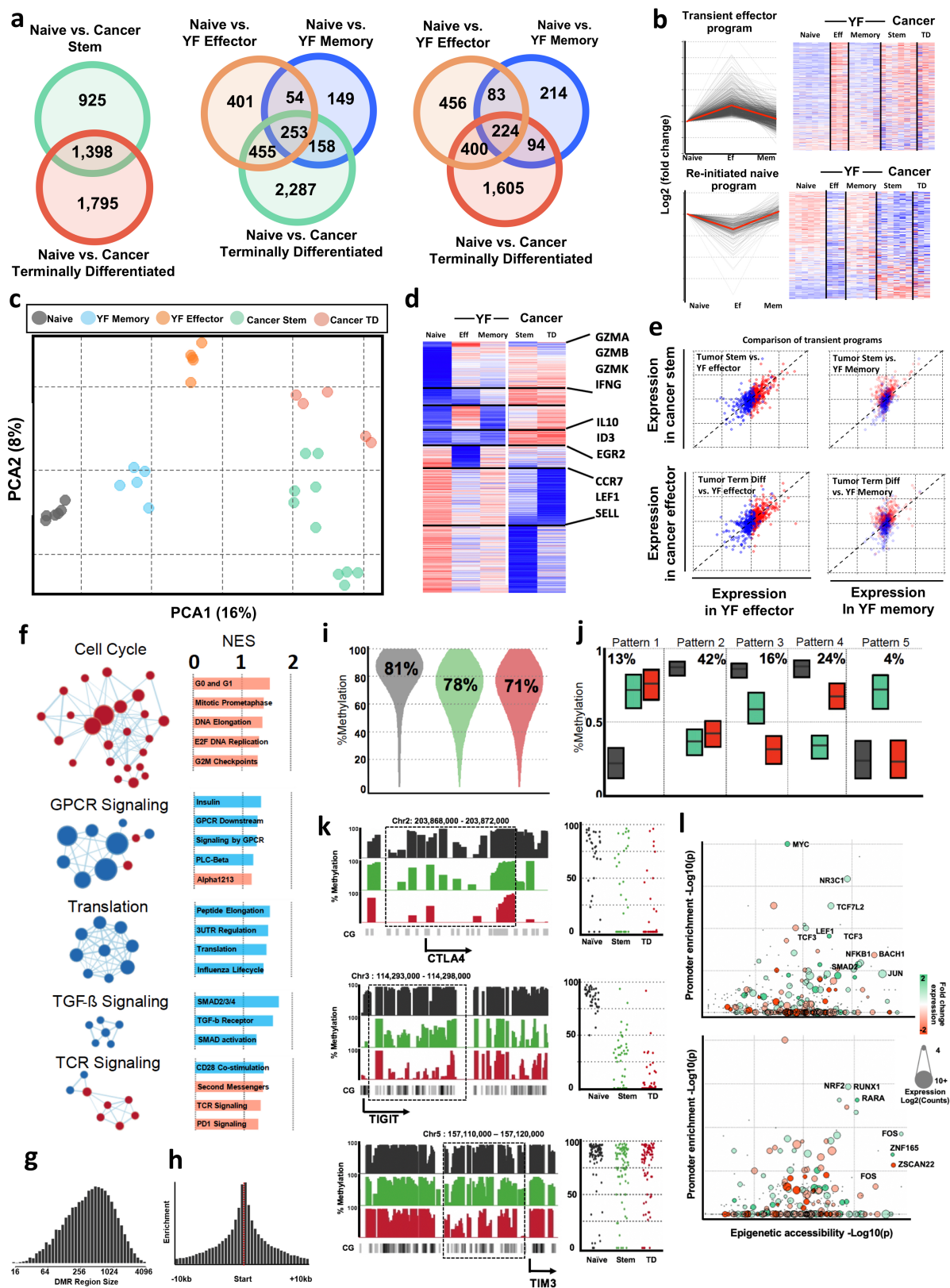
a, Flow cytometry gating scheme. FSC-A and FSC-H are used to select for singlets. Live (APC-Cy7 negative) CD3⁺ events are then selected from this population of singlets. Lymphocytes are selected from this live CD3⁺ population on the basis of FSC-A and SSC-A, and CD4⁺ and CD8⁺ T cell populations are selected from the lymphocyte population. **b**, Expression of various molecules by stem-like (green) and terminally differentiated (red) CD8 T cells in human tumours measured by flow cytometry. **c–e**, Expression of TCF1 (**c**), CD28 (**d**) and TIM3 (**e**) as measured by flow cytometry, by stem-like and terminally differentiated CD8 T cells isolated from patients with kidney cancer

(*n* = 6) and cultured in vitro for 3 days with 10 U of IL-2 and with (stimulated) or without (unstimulated) anti-CD3/CD28/CD2 bead stimulation at a 1:1 ratio. Median value is shown. **f**, Number of live stem-like and terminally differentiated intra-tumoral CD8⁺ T cells after 3 days of in vitro culture in IL-2 supplemented media. Live/dead staining was used to determine the proportion and number of live CD8 T cells by flow cytometry. **g**, Composition of the CD8 T cell compartment. In 60 human kidney cancer patients, proportion of CD8 T cells that are stem-like cells (PD-1⁺CD28⁺TIM3⁻) correlates with total T cell infiltration (%CD8 T cells of total cells), while proportion of terminally differentiated cells (PD-1⁺TIM3⁺) does not. **h**, Percentage of total CD8 T cells correlates with the percentage of total cells that are stem-like CD8 T cells.



Extended Data Fig. 4 | TCR sequencing analysis for stem-like and terminally differentiated CD8 T cells. **a**, Gating scheme for fluorescence activated cell sorting of cell populations for stem-like and terminally differentiated cell populations from human kidney tumours. Terminally differentiated cells¹ are PD-1-high and CD39⁺. Stem-like cells³ are PD-1⁺CD39⁻CD28⁺. **b**, Estimation of population overlap. PD-1 and CD39 expression by flow cytometry was modelled using a two-population Gaussian mixing model. The amount of each population falling within each sorting gate based on the relative proportions of the populations was determined and used to calculate whether TCRs found in both populations could be accounted for by contamination. **c**, Pre-sort flow cytometry plots for patients sorted for TCR sequencing. **d**, Ranking of stem-like (green) and terminally differentiated (red) TCR clones from most to 10th most dominant clone by percent of total TCR repertoire (log₁₀). **e**, Number of unique TCR clones detected in stem-like (green) and terminally differentiated (red) cell populations as a function of number of cells collected. **f**, Percentage

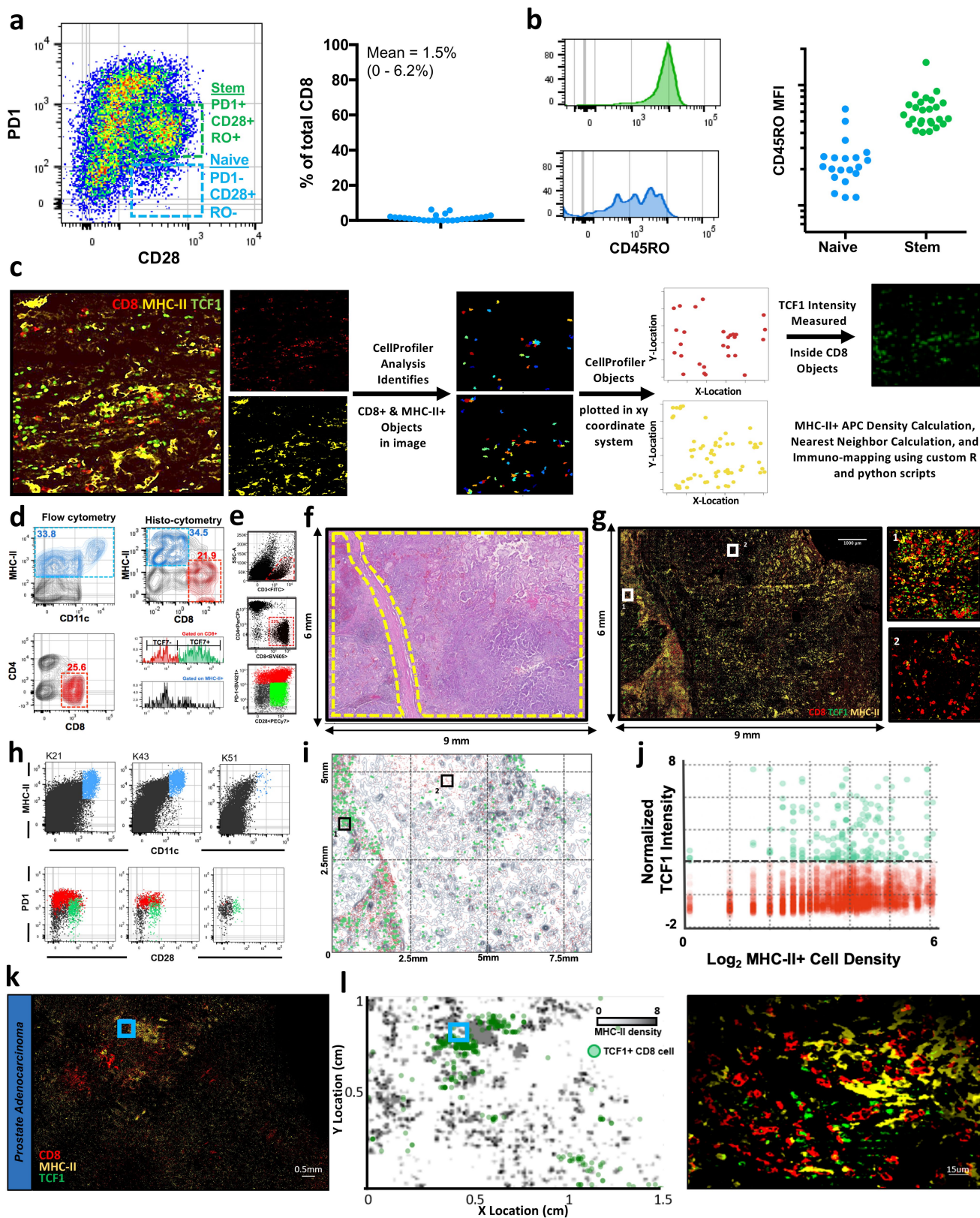
of overlap detected as a function of number of cells collected. **g**, Tumour samples were taken from two physically distant sites within the same tumour and stem-like and terminally differentiated cells were sorted from each and TCR sequenced. Venn diagrams illustrate unique TCRs found between stem-like populations in sites A and B, between terminally differentiated populations in sites A and B, and between location mismatched stem-like and terminally differentiated populations (for example, stem-like-A/terminally differentiated-B, stem-like-B/terminally differentiated-A), in addition to overlap between stem-like and terminally differentiated T cell populations within a single site. **h**, Table indicating the number of stem-like and terminally differentiated T cells collected, inferred purity of each population, percent overlap detected calculated by the number of TCRs detected in either sample divided by the total TCRs in both samples, and the power to detect >20% overlap (assuming 2,000 unique TCRs per sample) for each patient sample.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Transcriptional and epigenetic analysis of T cell subsets in tumours. a, Comparison of differentially expressed genes between human cancer and viral specific CD8 T cell subsets. RNA-seq from cancer subsets compared to RNA-seq data collected from yellow fever (YF) antigen specific CD8 T cells (GSE100745) during effector (14 days post-vaccination) and memory (4+ years post-vaccination) time points. The number of differentially expressed genes (DEG) versus naive CD8 T cells was determined using DESeq2. Venn diagrams show number of DEG shared or unique between viral and cancer subsets. Although the cancer subsets of T cells share many genes with the YF specific cells, there are also many distinct genes only expressed in cancer T cell subsets. **b,** DEGs were clustered using cluster affinity search technique (CAST). Clusters with greater than 5% of total genes are shown. Heat map shows z-score of averages from each group. **c,** Principal component analysis of T cell subsets from cancer and viral-specific CD8 T cells, performed on genes that were differentially expressed in any group versus naive cells. **d,** Comparison of cancer subsets to transient effector programs found in YF specific T cells. Previously we have identified transient gene expression signatures that are expressed in YF-specific effector cells, but return to a naive state after antigen is cleared. These genes not expressed in memory or naive cells are highly expressed in both cancer subsets suggesting a similarity to an effector cell. **e,** Pairwise comparison of transient effector program genes between effector and cancer subsets shows the relationship of this subset of genes re-initiated program (blue) and the transient effector program (red) compared between YF

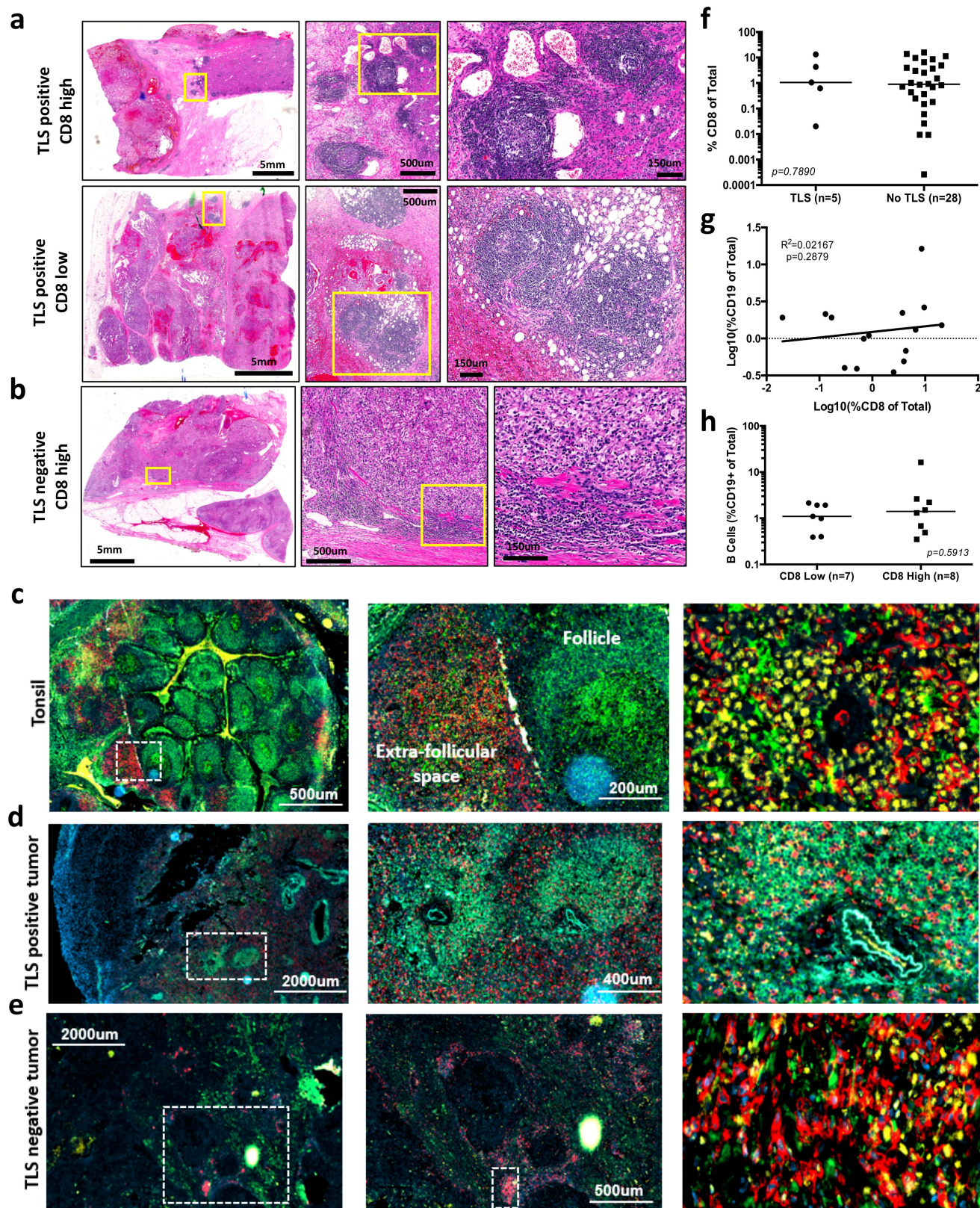
and cancer subsets. Dotted 45-degree line represents equal fold change versus a naive CD8 T cell in cancer and yellow fever cells. **f,** GSEA and network analysis of pathways associated with differentiation. Gene set enrichment performed with GSEA and visualized with Cytoscape. The most significant networks are shown. Red indicates enrichment of nodes in terminally differentiated T cells, while blue shows enrichment in stem-like T cells. **g,** Histogram shows the distribution of the continuous region size of DMRs. **h,** Histograms show the relative frequency of DMRs within 10kb of transcription start sites. **i,** Global changes in methylation. Violin plots show the distribution of total methylation within identified DMRs in naive, stem-like, and terminally differentiated cells. **j,** DMR patterns of differentiation. DMRs identified in Fig. 2d were clustered using CAST. Box plots show the interquartile range and mean of DMRs in each cluster by cell type **k,** Histograms show the total methylation from 0–100% in regions near important genes. Dot plots show the methylation of each CpG motif within highlighted regions of interest. **l,** Transcriptionally active transcription factors have over-represented binding in epigenetically modified regions of chromatin. Plots show the enrichment of transcription factor binding sites within differentially methylated regions in each cell type on the x-axis, and the y-axis shows the enrichment of transcription factor binding sites within the promoters of differentially expressed genes. Colour of dots represents the relative expression in stem-like (green) or terminally differentiated (red) cells, and the size of the dot is proportional to total expression of the transcription factor.



Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Quantitative immunofluorescence analysis of tumour immune infiltration. **a**, Flow cytometry data illustrating the number of naive cells present intra-tumorally. Left, representative patient. Right, summary data. **b**, Comparative amounts of CD45RO expression on naive and stem-like intra-tumoral CD8 T cells. **c**, Workflow for immunofluorescence imaging analysis and immuno-map creation. Single channel immunofluorescence images are imported into CellProfiler. CD8⁺ and MHC-II⁺ objects are identified in the respective channel images. The XY location of each CD8⁺ and MHC-II⁺ object is exported. The TCF1 staining intensity is measured inside the CD8⁺ objects. These parameters are used to calculate MHC-II⁺ density, measure the distance from each CD8⁺ object to its nearest MHC-II⁺ neighbour, and to finally create immuno-maps for immunofluorescence images. **d**, Histo-cytometric analysis of tumour infiltrating immune populations. Location and fluorescence intensity of CD8⁺ and MHC-II⁺ cells were determined using CellProfiler. After image compensation, CD8⁺ and MHC-II⁺ cells were gated. TCF1 intensity of each cell is shown on histograms for each population below. Comparison of flow cytometry data from the same patient sample is also shown. **e**, Patients with kidney cancer with high CD8 infiltration determined by

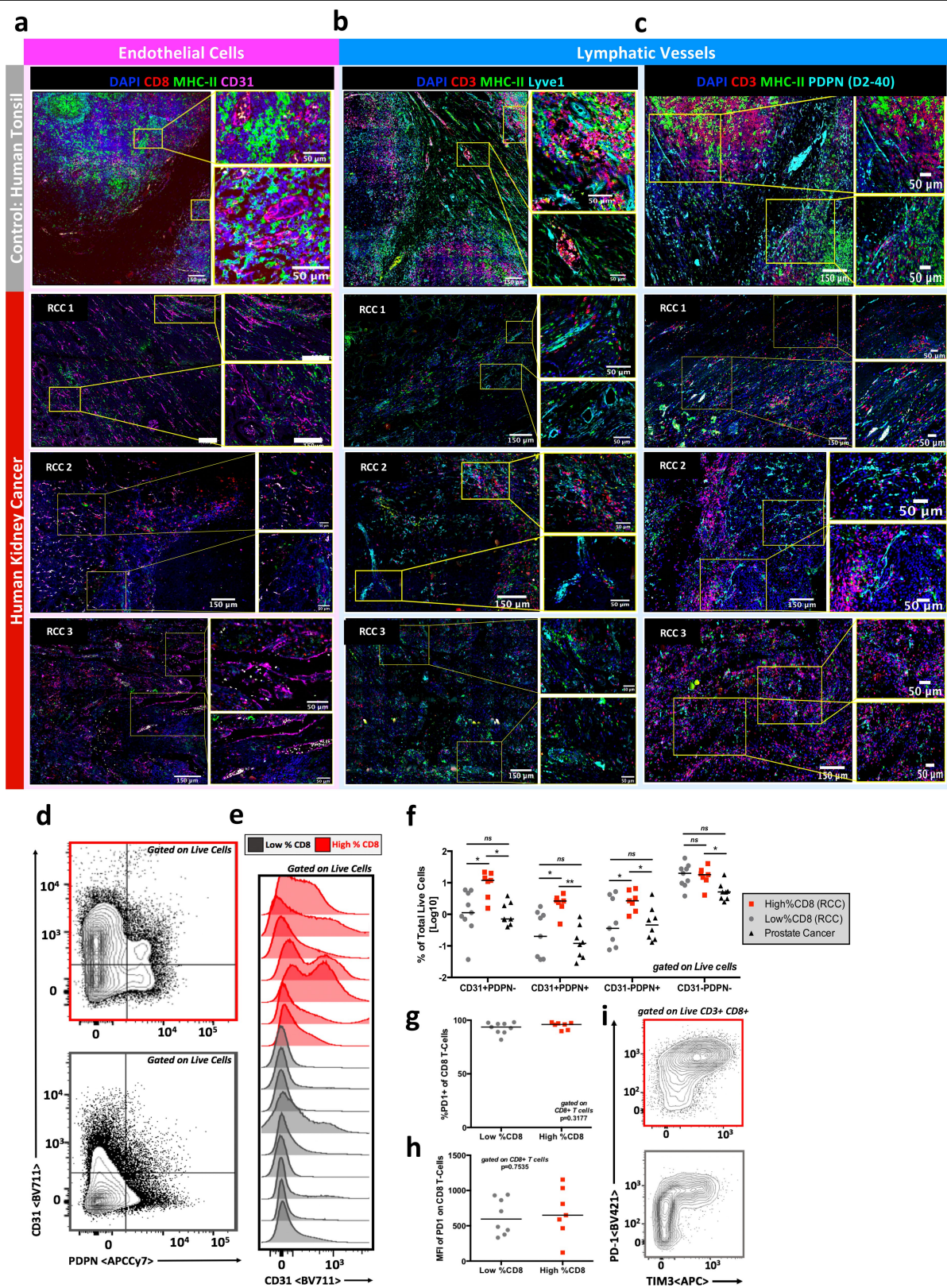
flow cytometry. Patients that were determined to have high CD8 infiltration by flow cytometry were selected for analysis by immunofluorescence. **f**, Haematoxylin and eosin stains of human kidney tumour. Selected slides from human kidney tumour shown in part **e**, to be highly infiltrated by T cells. Regions of tumour tissue are highlighted in yellow. **g**, Immunofluorescence imaging of kidney tumour. Selected tumours shown to be highly infiltrated by T cells. Tumour section was stained for MHC-II to identify antigen-presenting cells, and CD8 and TCF1 to identify stem-like and terminally differentiated CD8 T cell populations. Insets shows zoomed regions highlighted in the larger image. **h**, Dendritic cells populations, stem-like, and terminally differentiated CD8 T cells in three representative kidney cancer patients. **i**, Cellular spatial relationship map (middle) analysis and construction conducted as in Fig. 3e. **j**, CD8 expression of TCF1 preferentially occurs in dense APC zones. Amount of TCF1 expressed in each CD8 T cell graphed against the density of MHC-II around each T cell (MHC-II⁺ cells per 10,000 μm^2). **k**, **l**, TCF1⁺ CD8 T cells are localized near dense MHC-II regions in other cancers. Prostate and bladder tumours were imaged for CD8, MHCII and TCF1. Regions of dense MHC-II aggregates are shown in grey and the location of TCF1⁺ CD8 T cells in green (**l**).



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Comparison of tertiary lymphoid structures and antigen-presenting niches in kidney tumours. **a**, Haematoxylin and eosin slides highlighting tertiary lymphoid structures (TLS) in kidney tumours with high (top) and low (bottom) CD8 T cell infiltration. Yellow boxes highlight areas shown in zoomed insets. **b**, Haematoxylin and eosin slide showing dense immune infiltration in a tumour with high CD8 T cell infiltration but lacking presence of TLS. Yellow boxes highlight areas shown in zoomed insets. **c**, Immunofluorescence staining illustrating organizational structure of human tonsil. CD8 staining is shown in red, MHC-II in green, TCF1 in yellow and DAPI (nuclei) in blue. White box highlights zoomed area shown in inset. Follicle and extrafollicular space shown as labelled. T cell zone shown in rightmost panel. **d**, Immunofluorescence staining illustrating tumour TLS. CD8 staining is shown in red, MHC-II staining in green and DAPI staining of nuclei in blue. White box highlights zoomed area shown in inset. Follicle and extrafollicular

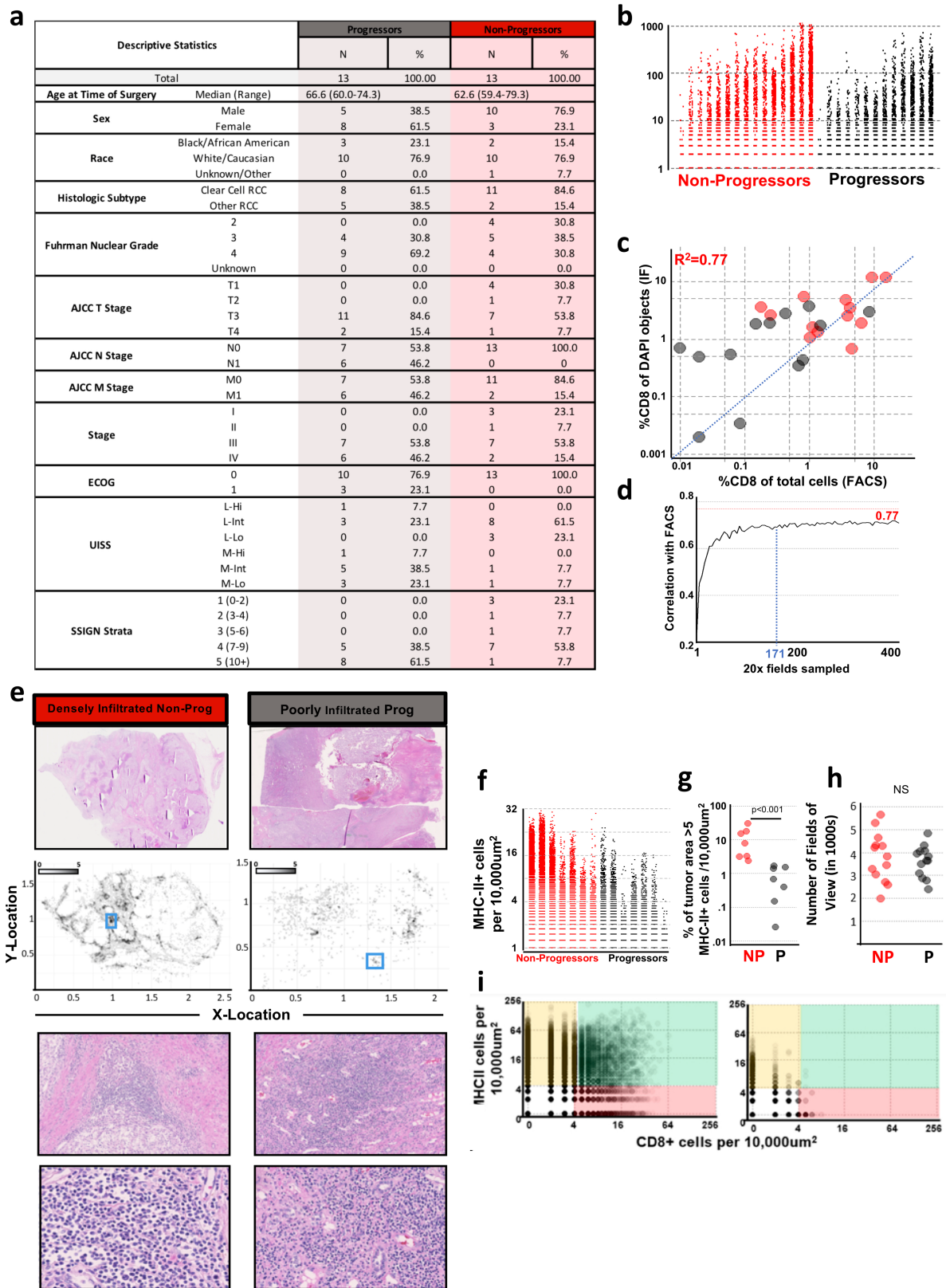
space shown as labelled. **e**, Immunofluorescence staining illustrating dense immune infiltration in TLS negative kidney tumour. CD8 staining is shown in red, MHC-II in green, TCF1 in yellow and DAPI in blue. White box highlights zoomed area shown in inset. Follicle and extrafollicular space shown as labelled. **f**, There is no significant difference in CD8 T cell infiltration between kidney tumours with and without TLS. CD8 T cell infiltration measured by flow cytometry and shown as percentage of CD8⁺ of total cells. Statistical analysis resultant from Mann–Whitney test is shown. **g**, Lack of correlation between proportion of CD8 T cells and CD19⁺ B cells in tumours. Linear regression results $P=0.6006$ with $R^2=0.02167$. **h**, B cell infiltration between tumours with high or low CD8 T cell infiltration was not significantly different. B cell infiltration is shown as the percentage of CD19⁺ B cells of total cells. Statistical analysis resultant from Mann–Whitney test is shown. Median value shown in **f** and **h**.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Highly infiltrated kidney tumours are well vascularized and contain lymphatic vessels. **a**, Immunofluorescence staining of human tonsil and highly T cell infiltrated human kidney tumours showing tissue vascularization. Formalin-fixed paraffin embedded tissue was stained for CD8 (T cells), MHC-II (antigen-presenting cells), CD31 (endothelial cells) and DAPI (nuclei). **b, c**, Immunofluorescence staining of human tonsil and highly T cell infiltrated kidney tumours showing presence of lymphatics via Lyve1 (**b**) and Podoplanin/D2-40 (**c**). Formalin-fixed paraffin embedded tissue was stained for CD3 (T cells), MHC-II (antigen presenting cells), Lyve 1 or Podoplanin/D2-40 (lymphatics) and DAPI (nuclei). **d**, Flow cytometry analysis shows tumour vascularization in highly (red) and poorly (grey) infiltrated kidney tumour. Tumours were stained using antibodies listed in Supplementary Table 2, collected on a Becton Dickinson LSR-II, and analysed using FlowJo. **e**, Histogram of flow cytometry analysis showing increased CD31

staining in highly T cell-infiltrated kidney tumours (red) as compared to poorly infiltrated tumours (grey). Analysis completed as described in **d, f**. Summary data of flow cytometry analysis showing differences in vascularization between highly (red) and poorly (grey) T cell infiltrated kidney tumours and prostate tumours (black). Analysis completed as described in **d, g, h**. Tumour-infiltrating T cells are PD-1⁺. Flow cytometry analysis showing T cells infiltrating kidney tumours are PD-1⁺, suggesting the cells are not naive and present due to blood contamination and showing that the MFI of PD-1 on tumour-infiltrating T cells is not significantly different between highly (red) and poorly (grey) infiltrated tumours. **i**, Representative flow cytometry plots showing PD-1 and TIM3 expression on tumour infiltrated T cells in highly (red) and poorly (grey) infiltrated tumours. Populations shown are gated on live, CD3⁺CD8⁺ cells. Median value shown in **f-h**.

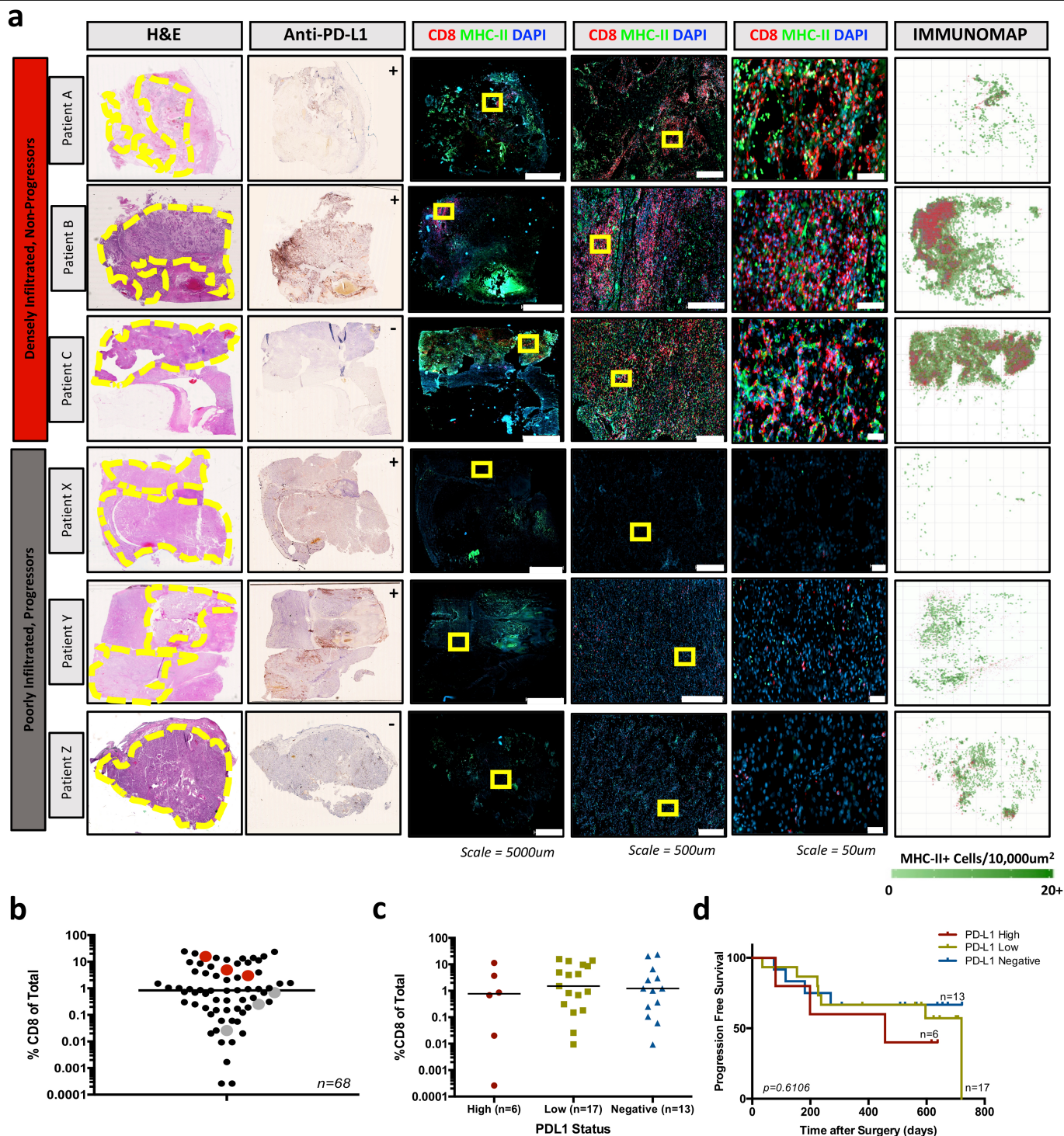


Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Descriptive statistics and quantitative immunofluorescence analyses of human kidney tumours.

a, Descriptive table enumerating patient characteristics of patients with kidney cancer, with and without progressive disease. **b**, Comparison of the number of CD8⁺ cells per 300 $\mu\text{m} \times 300 \mu\text{m}$ field in patients with and without progressive disease. The number of CD8⁺ cells per 300 $\mu\text{m} \times 300 \mu\text{m}$ field were enumerated using the methods outlined in Extended Data Fig. 6. **c**, The correlation between enumeration of CD8 T cells by flow cytometry and by immunofluorescence. On the *x* axis, CD8 T cells are measured as a proportion of total cells. On the *y* axis, CD8 T cells are measured as a proportion of total DAPI objects detected in the tumour section. **d**, Estimated number of 20 \times fields of view necessary to obtain an accurate assessment of level of CD8 T cell infiltration is 171 fields of view. Increasing number of random fields of view were sampled from images and the percent of cells that were CD8 positive by IF correlated to FACS from the corresponding sample. **e**, Histological comparison of patients with kidney cancer shown in Fig. 4 – a patient with kidney cancer with dense T cell

infiltration and no disease progression (red, left) and a patient with kidney cancer with poor T cell infiltration and progressive disease (grey, right). **f**, Comparison of the number of MHC-II⁺ cells per 300 $\mu\text{m} \times 300 \mu\text{m}$ field in stage III (T3N0M0) patients with and without progressive disease. The number of MHC-II⁺ cells per 300 $\mu\text{m} \times 300 \mu\text{m}$ field were enumerated using the methods outlined in Extended Data Fig. 6. **g**, Comparison of the proportion of tumour area with greater than 5 MHC-II⁺ cells per 10,000 μm^2 between stage III (T3N0M0) patients with and without progressive disease. Statistical analysis resultant from Mann–Whitney test is shown. **h**, No significant difference in number of fields of view sampled between patients with and without progressive disease was detected. **i**, Density of MHC-II⁺ APCs and CD8 T cells in densely (left) or poorly (right) infiltrated kidney tumours. *x*-axis shows the number of CD8⁺ cells per 10,000 μm^2 . *y*-axis shows the number of MHC-II⁺ cells per 10,000 μm^2 . Regions of predominantly MHC-II⁺ cells are highlighted in yellow, regions of predominantly CD8⁺ cells in red, and regions of shared MHC-II⁺ cells and CD8⁺ cells in green.



Extended Data Fig. 10 | Comparison of densely and poorly infiltrated kidney tumours by PDL-1 staining and by quantitative immunofluorescence. a, Representative patients with densely infiltrated and poorly infiltrated kidney tumours whose disease has not progressed or has progressed, respectively. Whole-slide scans are shown for haematoxylin and eosin, anti-PD-L1, and immunofluorescence (CD8, MHC-II, DAPI) stains, with zoomed insets of immunofluorescence data. Yellow circles highlight the location of tumour tissue on the haematoxylin and eosin slide. Yellow boxes highlight the areas shown in the zoomed insets of immunofluorescence images. Immunofluorescence data are quantitatively analysed and mapped to show the density of MHC-II⁺ cells and the XY location of CD8⁺ T cells in the rightmost panel. Anti-PD-L1 scans are marked as ++ (positive-high), + (positive-low),

or - (negative), as scored by board-certified pathologists. **b,** Patients in **a** are highlighted in red (highly infiltrated, non-progressors) and grey (poorly infiltrated, progressors) to show the percentage of CD8 T cell infiltration by flow cytometry. **c,** PD-L1 staining was scored by board-certified pathologists as positive-high, positive-low and negative. There is no significant difference between the percent CD8 T cell infiltration amongst these categories by ANOVA with Holm-Sidak correction. Median value shown. **d,** Progression free survival for patients with positive-high (PD-L1 high), positive-low (PD-L1 low), and negative (PD-L1 negative) kidney tumours. There was no significant difference in progression-free survival between the groups by Mantel-Cox log rank test ($P = 0.6106$) or by log rank test for trend ($P = 0.3374$).

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Data was collected using FACSDiva for flow cytometry and Leica LASX for IF.

Data analysis

FACS data was analyzed with FlowJo V10, IF data was analyzed using CellProfiler V3 and custom R scripts, RNAseq and DNA methylation data was analyzed using DESeq2 and custom R scripts.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw fastq files and associated RNA and WGBS sequencing analysis have been uploaded to the NCBI Gene Expression Omnibus (GEO) database. Other relevant data are available from the corresponding author upon reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample sizes were determined prior to study
Data exclusions	4 RNAseq samples were discarded because of very low cell recovery and subsequent poor sequencing data. 5 Immunofluorescence samples were excluded because of extensive artifact that made quantification of cells obviously inaccurate.
Replication	All data presented include the patients analyzed. No experiments are excluded.
Randomization	No Randomization was used in this study
Blinding	No Blinding was done in this study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used	Provided in methods section
Validation	Antibodies are validated as described on manufacturers product information

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Patients used in this study were sequentially collected between June 2014 and January 2018
Recruitment	Patients undergoing surgery in the department of Urology at Emory University were consented to collect bio-specimens during their care.
Ethics oversight	Emory IRB

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☐ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Described in methods sections
Instrument	The data was collected on a Becton Dickinson LSR-II cytometer or Becton Dickinson Canto-II cytometer
Software	Data was analyzed using FlowJo V10
Cell population abundance	When greater than 5000 cells were collected we performed post-sort purity. In all cases this was greater than 95%. For analysis of markers on T-cell populations, we required at least 1000 cells.
Gating strategy	Gating strategies are shown in extended data figures
<input checked="" type="checkbox"/> Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.	

Targeting REGNASE-1 programs long-lived effector T cells for cancer therapy

<https://doi.org/10.1038/s41586-019-1821-z>

Received: 5 February 2019

Accepted: 7 November 2019

Published online: 11 December 2019

Jun Wei^{1,5}, Lingyun Long^{1,5}, Wenting Zheng², Yogesh Dhungana¹, Seon Ah Lim¹, Cliff Guy¹, Yanyan Wang¹, Yong-Dong Wang³, Chenxi Qian^{1,3}, Beisi Xu³, Anil KC¹, Jordy Saravia¹, Hongling Huang¹, Jiyang Yu³, John G. Doench⁴, Terrence L. Geiger² & Hongbo Chi^{1*}

Adoptive cell therapy represents a new paradigm in cancer immunotherapy, but it can be limited by the poor persistence and function of transferred T cells¹. Here we use an in vivo pooled CRISPR–Cas9 mutagenesis screening approach to demonstrate that, by targeting REGNASE-1, CD8⁺ T cells are reprogrammed to long-lived effector cells with extensive accumulation, better persistence and robust effector function in tumours. REGNASE-1-deficient CD8⁺ T cells show markedly improved therapeutic efficacy against mouse models of melanoma and leukaemia. By using a secondary genome-scale CRISPR–Cas9 screening, we identify BATF as the key target of REGNASE-1 and as a rheostat that shapes antitumour responses. Loss of BATF suppresses the increased accumulation and mitochondrial fitness of REGNASE-1-deficient CD8⁺ T cells. By contrast, the targeting of additional signalling factors—including PTPN2 and SOCS1—improves the therapeutic efficacy of REGNASE-1-deficient CD8⁺ T cells. Our findings suggest that T cell persistence and effector function can be coordinated in tumour immunity and point to avenues for improving the efficacy of adoptive cell therapy for cancer.

Adoptive cell therapy (ACT), including the use of T cells engineered to express chimeric antigen receptors (CARs), has produced unprecedented clinical outcomes for cancer immunotherapy. However, the therapeutic efficacy of ACT—especially in solid tumours—is often limited by the poor in vivo accumulation, persistence and function of adoptively transferred T cells¹. Paradoxically, terminal effector CD8⁺ T cells have been shown to have reduced antitumour efficacy and exhibit poor in vivo persistence². How T cell fate decisions are regulated in the tumour microenvironment (TME) remains poorly understood.

Here, through an in vivo pooled CRISPR–Cas9 mutagenesis screening of metabolism-associated factors, we identify REGNASE-1 as a major negative regulator of antitumour responses. REGNASE-1-deficient CD8⁺ T cells are reprogrammed in the TME to long-lived effector cells by enhancing BATF function and mitochondrial metabolism, thereby improving ACT for cancer.

Screen for metabolic regulators of ACT

T cell longevity and function in cancer immunotherapy have previously been proposed to closely correlate with cell metabolic fitness³, although the underlying molecular mechanisms are unclear. To systematically investigate the roles of metabolism-associated factors in T-cell-mediated antitumour immunity, we developed a pooled CRISPR–Cas9 mutagenesis screening approach in an ACT model (Fig. 1a), using CD8⁺ T cells that express the OT-I T cell receptor (TCR) and Cas9 as well as mice inoculated with B16 melanoma cells that express the cognate antigen (B16 Ova). We developed two lentiviral sub-libraries of single-guide

(sg) RNAs (six sgRNAs per gene) targeting 3,017 metabolic enzymes, small molecule transporters and metabolism-related transcriptional regulators (Supplementary Table 1). Seven days after adoptive transfer, sgRNA-transduced OT-I cells in tumour-infiltrating lymphocytes were examined for library representation. A total of 218 genes were significantly depleted (Fig. 1b, Supplementary Table 2), including *Txnrd1*⁴, *Ldha*⁵, *Fth1*⁶ and *Foxo1*⁷, which are known regulators of cell survival and expansion. Notably, *Zc3h12a* (also known as *Regnase-1*, which encodes REGNASE-1) was the most highly enriched gene (Fig. 1b), which suggests that REGNASE-1 could be a major negative regulator of antitumour responses. REGNASE-1 is known to have RNase activity and to regulate activation of immune cells^{8,9}, but the function of REGNASE-1 in tumour immunity is currently unclear.

To validate our findings, we developed an in vivo dual transfer system to compare OT-I cells that were transduced with sgRNA vectors that express distinct fluorescent proteins in the same tumour-bearing host (Extended Data Fig. 1a, b), without there being noticeable effects of different fluorescent proteins per se (top panels in Extended Data Fig. 1c, d). We tested OT-I cells transduced with two different sgRNAs that target *Regnase-1*, and found that the relative proportion of REGNASE-1-null OT-I cells was markedly increased in both the spleen and tumour (Extended Data Fig. 1c–e). Imaging analysis identified significantly more REGNASE-1-null OT-I cells than wild-type control cells within tumours (Fig. 1c). Analysis of the targeting efficacy of guides revealed efficient disruption of *Regnase-1* (Extended Data Fig. 1f). Next, we examined the persistence of REGNASE-1-null OT-I cells at days 7, 14 and 21 after transfer. Whereas wild-type OT-I cells declined over time,

¹Department of Immunology, St Jude Children's Research Hospital, Memphis, TN, USA. ²Department of Pathology, St Jude Children's Research Hospital, Memphis, TN, USA. ³Department of Computational Biology, St Jude Children's Research Hospital, Memphis, TN, USA. ⁴Broad Institute of Harvard and MIT, Cambridge, MA, USA. ⁵These authors contributed equally: Jun Wei, Lingyun Long. *e-mail: hongbo.chi@stjude.org

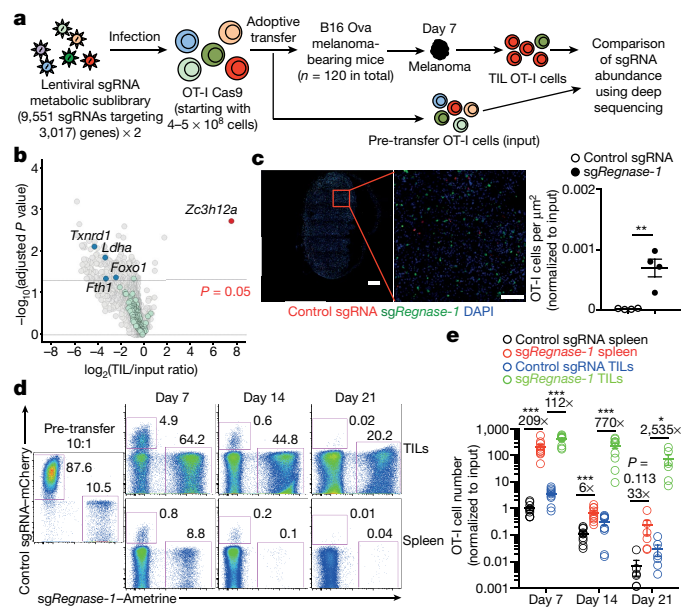


Fig. 1 | In vivo CRISPR–Cas9 screening identifies REGNASE-1 as a major negative regulator of the antitumour responses of CD8⁺ T cells. **a**, Diagram of CRISPR screening for metabolic regulators of ACT. TIL, tumour-infiltrating lymphocyte. **b**, Scatter plot of the enrichment of candidates ($n = 6$ sgRNAs per gene) with the most extensively enriched (red) and selectively depleted (blue) genes, as well as dummy genes (green; generated by random combinations of 6 out of 1,000 non-targeting control sgRNAs per dummy gene) highlighted. **c**, Representative images (left) and quantification of the relative OT-I cell number per area (μm^2) normalized to input (right) in the tumour section ($n = 4$). OT-I cells transduced with control sgRNA (red) and sgRNA against *Regnase-1* (sg*Regnase-1*, green) were mixed at a 10:1 ratio and transferred into tumour-bearing mice, and analysed at day 7. Scale bars, 500 μm . **d**, **e**, OT-I cells transduced with control sgRNA or sgRNA against *Regnase-1* were mixed at a 10:1 ratio and transferred into tumour-bearing mice, followed by analysis of the proportion of OT-I cells in total CD8 α^+ cells (**d**) and quantification of normalized OT-I cell number relative to input (**e**) at day 7 ($n = 10$), day 14 ($n = 10$) and day 21 ($n = 6$). Cell numbers in the tumour are indicated per gram of tissue. Mean \pm s.e.m. (**c**, **e**). * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Two-tailed paired Student's *t*-test followed by Bonferroni correction (**b**) or two-tailed unpaired Student's *t*-test (**c**, **e**). Data are representative of two (**c**, **d**), or pooled from two (**e**), independent experiments.

REGNASE-1-null cells had markedly better persistence—especially in the tumour and at later time points (Fig. 1d, e). Therefore, the loss of REGNASE-1 endows tumour-specific CD8⁺ T cells with greatly improved accumulation and persistence, preferentially in the tumour.

Loss of REGNASE-1 improves ACT efficacy

We assessed the efficacy of REGNASE-1-null CD8⁺ T cells in ACT that targets a range of different tumours. In the B16 Ova model of melanoma, REGNASE-1-null OT-I cells showed much stronger antitumour effects than did wild-type cells, as evidenced by markedly inhibited tumour growth and the increased survival of melanoma-bearing mice (Fig. 2a, b). Similar results were observed in T cells that express the pme-1 TCR (which recognize the endogenous melanoma antigen gp100) when transferred into mice bearing B16 F10 melanoma (Fig. 2c, d). To assess the efficacy of CAR T cells against leukaemia, we used T cells that express CARs targeting human CD19 in combination with BCR–ABL1⁺ progenitor acute lymphoblastic leukaemia (Philadelphia-chromosome (Ph)⁺ B-ALL) cells¹⁰ that express human CD19 (human CD19–Ph⁺ B-ALL cells). REGNASE-1-null CAR T cells showed a therapeutic efficacy that was much stronger than that of wild-type cells, as indicated by mouse survival (Fig. 2e) and tumour burden analyses

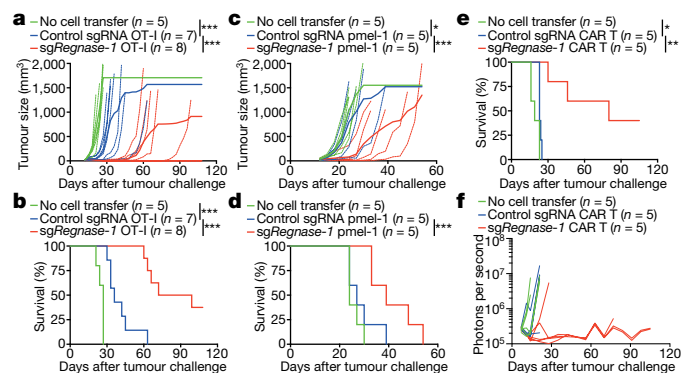


Fig. 2 | Deletion of REGNASE-1 enhances the efficacy of ACT against solid and blood cancers. **a–f**, OT-I (**a**, **b**), pmel-1 (**c**, **d**) or CD8⁺ CAR T (**e**, **f**) cells (5×10^6) transduced with non-targeting control sgRNA or sgRNA against *Regnase-1* were transferred into mice at day 12 after engraftment of B16 Ova (**a**, **b**) or B16 F10 (**c**, **d**) melanoma, or at day 7 after Ph⁺ B-ALL cell engraftment (**e**, **f**), followed by analyses of tumour size (**a**, **c**), mouse survival (**b**, **d**, **e**) and tumour burden via xenogen imaging of bioluminescent signal intensities (**f**). Non-treatment control mice received no transfer of T cells. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Two-way analysis of variance (ANOVA) (**a**, **c**) or log-rank (Mantel–Cox) test (**b**, **d**, **e**). Data are representative of two (**a**, **b**, **e**, **f**) or four (**c**, **d**) independent experiments.

(Fig. 2f). Collectively, REGNASE-1 deletion markedly enhances the efficacy of ACT against both solid and blood cancers.

REGNASE-1 loss reprograms T cells in TME

To address cell-intrinsic effects mediated by REGNASE-1, we performed RNA sequencing (RNA-seq) of REGNASE-1-null and wild-type OT-I cells isolated from the in vivo dual transfer system. Gene set enrichment analysis (GSEA) using gene modules associated with different functional states of tumour-infiltrating CD8⁺ T cells¹¹ revealed that tumour-infiltrating REGNASE-1-null cells were enriched with the naive or memory module (Fig. 3a). Gene targets that were repressed by REGNASE-1 were also enriched in memory-like CD8⁺ T cells in chronic infection^{12,13} (Extended Data Fig. 2a, b). Accordingly, tumour-infiltrating REGNASE-1-null cells had increased expression of TCF-1 (a transcription factor that is associated with naive or memory T cells¹⁴) (Fig. 3b, Extended Data Fig. 2c) and *Lef1*, *Bach2*, *Tcf7* (which encodes TCF-1), *Foxp1*, *Bcl6* and *Fosb*^{14–17}, but had lower expression of *Irf2*, *Irf4* and *Hmgb2*^{18,19} (Extended Data Fig. 2d–f). We next performed assay for transposase-accessible chromatin using sequencing (ATAC-seq)²⁰ to measure the chromatin accessibility of tumour-infiltrating REGNASE-1-null and wild-type cells. Motif searches on accessible regions identified an enrichment of TCF-1, BACH2 and BCL6—but downregulated IRF4 motifs—in REGNASE-1-null cells (Extended Data Fig. 2g, h). Thus, REGNASE-1-null CD8⁺ T cells are reprogrammed in the TME with enhanced gene-expression programs associated with naive or memory cells.

Transcriptional profiling revealed marked differences between tumour-infiltrating and peripheral REGNASE-1-null cells (Extended Data Fig. 2i). Unlike the enrichment of the naive or memory module in REGNASE-1-null cells in tumours (Fig. 3a), but consistent with previous reports that describe the negative role of REGNASE-1 in T cell activation under homeostasis^{8,9}, peripheral REGNASE-1-null cells were enriched with the activation-associated—but not with the naive or memory—module (Extended Data Fig. 2j), and had reduced expression of TCF-1 (Extended Data Fig. 2k). Given the TME-specific phenotypes of REGNASE-1-null cells, we assessed the regulation of REGNASE-1 and found lower expression of REGNASE-1 in tumour-infiltrating than in peripheral OT-I cells (Extended Data Fig. 3a). Additionally, gene targets repressed by REGNASE-1 were increased in tumour-infiltrating

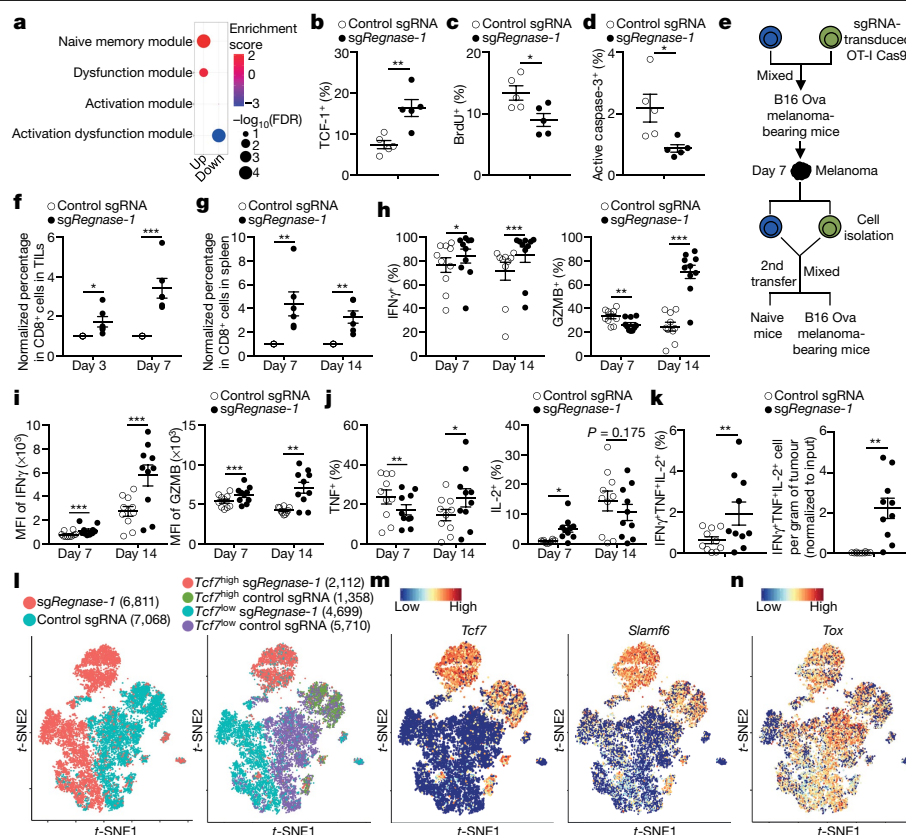


Fig. 3 | Deletion of REGNASE-1 reprograms tumour-infiltrating CD8⁺ T cells to long-lived effector cells. **a**, GSEA enrichment plots of RNA-seq analysis of OT-I cells, transduced with sgRNA against *Regnase-1* ($n = 5$) versus non-targeting control sgRNA, and isolated from tumour-infiltrating lymphocytes from the dual transfer system, using gene sets of the activation states of tumour-infiltrating CD8⁺ T cells¹¹. **b–d**, Tumour-infiltrating OT-I cells transduced with sgRNA, from the dual transfer system ($n = 5$), were analysed at day 7 (**b**) and day 14 (**c, d**) for the quantification of frequencies of TCF-1⁺ (**b**), BrdU⁺ (**c**) and active caspase-3⁺ (**d**) cells. **e–g**, Diagram of in vivo persistence assay (**e**): sgRNA-transduced OT-I cells were isolated from tumour-infiltrating lymphocytes, mixed at a 1:1 ratio (1×10^5 each) and transferred into tumour-bearing hosts (**f**) or naive mice (**g**). Quantification of the normalized OT-I cell frequency in tumour-infiltrating lymphocytes of tumour-bearing hosts ($n = 6$) (**f**) or in the spleen of naive hosts ($n = 6$) (**g**). **h–k**, Tumour-infiltrating OT-I cells transduced with sgRNA, from the dual transfer system, were analysed at day 7

($n = 10$) and day 14 ($n = 10$) for the quantification of frequencies of IFN γ ⁺ cells (**h**, left), GZMB⁺ cells (**h**, right), TNF⁺ cells (**j**, left), IL-2⁺ cells (**j**, right) and polyfunctional IFN γ ⁺TNF⁺IL-2⁺ cells (**k**, left) in OT-I cells, and mean fluorescence intensity (MFI) of IFN γ and GZMB in IFN γ ⁺ and GZMB⁺ cells, respectively (**i**), and cell number (normalized to input) per gram of tumour (**k**, right) of polyfunctional IFN γ ⁺TNF⁺IL-2⁺ OT-I cells. **l–n**, scRNA-seq analysis of tumour-infiltrating OT-I cells transduced with sgRNA, isolated from the dual transfer system at day 7. **l**, Distributed stochastic neighbour embedding (t-SNE) visualization of OT-I cells indicating genotypes (**l**, left), *Tcf7*^{high} and *Tcf7*^{low} cells (**l**, right), and *Tcf7* (**m**, left), *Slamf6* (**m**, right) and *Tox* (**n**) gene expression in individual cells. Mean \pm s.e.m. (**b–d, f–k**). * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Kolmogorov–Smirnov test followed by Benjamini–Hochberg correction (**a**), two-tailed unpaired Student's *t*-test (**b–d, f, g**) or two-tailed paired Student's *t*-test in (**h–k**). Data are representative of three (**b**), or pooled from two (**c, d, f–k**), independent experiments.

cells (Extended Data Fig. 3b), indicative of dampened REGNASE-1 activity. Moreover, stimulation with TCR—and, to a lesser extent, IL-2 or IL-21—induced REGNASE-1 cleavage⁹ (Extended Data Fig. 3c). Antigen recognition was crucial in driving the CD8⁺ T cell accumulation in tumour-infiltrating lymphocytes upon deletion of REGNASE-1, as indicated by reduced REGNASE-1-null OT-I cells in mice that bear B16 F10 melanoma (without the cognate antigen) compared to B16 Ova melanoma (Extended Data Fig. 3d). Antigen stimulation was also required for REGNASE-1-null cells to acquire increased TCF-1 expression (Extended Data Fig. 3e). By contrast, hypoxia did not alter the expression of REGNASE-1 or immune markers (Extended Data Fig. 3f, g). Thus, REGNASE-1-null CD8⁺ T cells undergo specific reprogramming in the TME in a process downstream of tumour antigen stimulation.

We next determined the cellular homeostasis of REGNASE-1-null cells. GSEA revealed that cell-cycling-associated hallmarks were the top downregulated pathways in tumour-infiltrating REGNASE-1-null cells (Extended Data Fig. 4a, b). Accordingly, these cells had reduced BrdU and Ki-67 staining at day 14 after adoptive transfer (Fig. 3c, Extended Data Fig. 4c), albeit not at day 7 (Extended Data Fig. 4d, e).

Also, tumour-infiltrating REGNASE-1-null cells had reduced levels of active caspase-3 (Fig. 3d, Extended Data Fig. 4f) and DNA damage biomarker (Extended Data Fig. 4g). Therefore, tumour-infiltrating REGNASE-1-null cells are less proliferative after effector expansion and show better survival than wild-type cells. By contrast (but consistent with the increased activation signatures; Extended Data Fig. 2j), peripheral REGNASE-1-null cells were enriched with signatures associated with cell cycling and apoptosis (Extended Data Fig. 4h), which was validated by increased BrdU and active caspase-3 staining (Extended Data Fig. 4i, j). These results further support the TME-specific phenotypes of REGNASE-1-null CD8⁺ T cells. To test the effect on in vivo persistence, we isolated wild-type and REGNASE-1-null OT-I cells from tumour-infiltrating lymphocytes and cotransferred them into tumour-bearing or naive hosts (Fig. 3e). REGNASE-1-null cells showed increased accumulation in tumour sites (Fig. 3f) as well as in the spleen of naive recipients (Fig. 3g). These analyses collectively indicate that tumour-infiltrating REGNASE-1-null CD8⁺ T cells are characterized by in vivo quiescence and survival, with better persistence than wild-type CD8⁺ T cells in response to both antigenic and homeostatic signals.

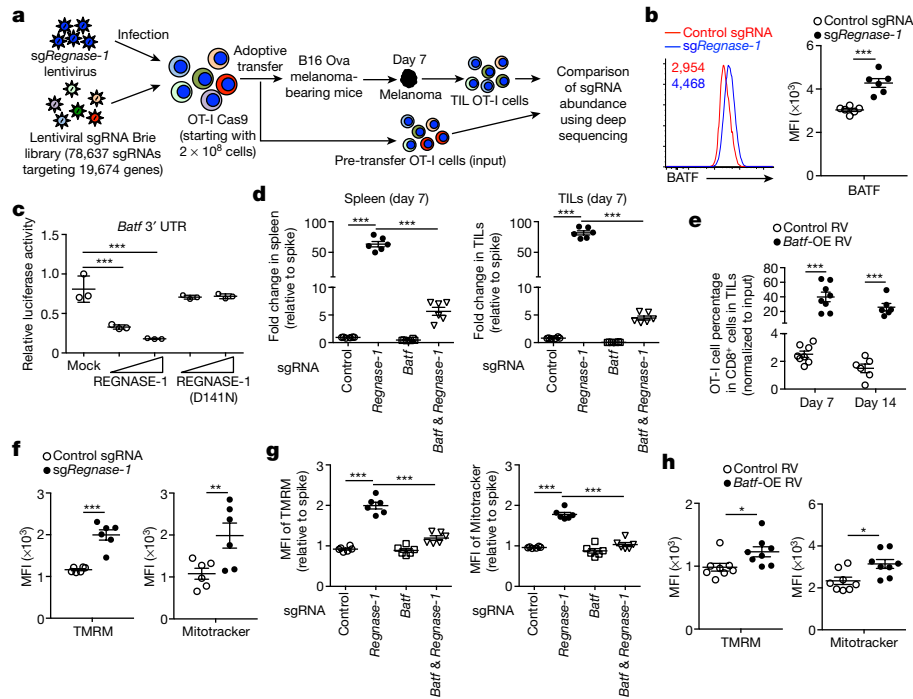


Fig. 4 | BATF is a key REGNASE-1 functional target for mediating mitochondrial fitness and effector responses. **a**, Diagram of secondary genome-scale CRISPR screening. **b**, Tumour-infiltrating OT-I cells transduced with sgRNA, from the dual transfer system ($n = 6$), were analysed at day 7 for BATF expression (left) and quantification of BATF MFI (right). **c**, Luciferase activity of HEK293T cells after transfection with *Batf* mRNA 3' UTR reporter, together with control (mock), wild-type or REGNASE-1(D141N)-expressing plasmid ($n = 3$). **d, e**, In vivo accumulation of OT-I cells transduced with an individual sgRNA or with two sgRNAs (**d**) or *Batf*-overexpressing (OE) retrovirus (RV) (**e**) in the dual transfer system ($n = 6$). The OT-I cell percentage in CD8 α^+ cells was normalized to co-transferred spike cells transduced with the

non-targeting control sgRNA (**d**). **f–h**, Tumour-infiltrating OT-I cells transduced with an individual sgRNA (**f**; $n = 6$), individual or two sgRNAs (**g**; $n = 6$) or *Batf*-overexpressing retrovirus (**h**; $n = 8$) from the dual transfer system were analysed at day 7 for the quantification of the MFI of tetramethylrhodamine, methyl ester (TMRM) (left) and Mitotracker (right). The MFIs of TMRM and Mitotracker were normalized to those of co-transferred spike cells transduced with control sgRNA (**g**). Mean \pm s.e.m. (**b, d–h**) or mean \pm s.d. (**c**). * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Two-tailed unpaired Student's *t*-test (**b, e, h**), one-way ANOVA (**c, d, g**) or two-tailed paired Student's *t*-test (**f**). Data are representative of two (**c**), or pooled from two (**b, d–h**), independent experiments.

Although in tumours REGNASE-1-null OT-I cells acquired programs associated with naive or memory cells, these cells had higher expression of many activation-associated markers (Extended Data Fig. 5a), retained an effector surface phenotype (CD44 $^+$ CD62L $^+$) (Extended Data Fig. 5b), and expressed more IFN γ and granzyme B (GZMB) (Fig. 3h, i, Extended Data Fig. 5c, d). Additionally, these cells had similar or enhanced capacities to produce TNF and IL-2 (Fig. 3j, Extended Data Fig. 5e, f), and contained an increased number of IFN γ^+ TNF $^+$ IL-2 $^+$ polyfunctional cells (Fig. 3k). Thus, although tumour-infiltrating CD8 $^+$ T cells that lack REGNASE-1 acquire better persistence and a survival advantage, they retain potent effector function.

We used single-cell (sc)RNA-seq²¹ to investigate the heterogeneity of tumour-infiltrating lymphocytes isolated from the in vivo dual transfer system. REGNASE-1-null OT-I cells had patterns that were distinct from those of wild-type cells, including an increased proportion of *Tcf7*^{high} cells (Fig. 3l, m). In both genotypes, *Tcf7*^{high} cells expressed *Tox*^{22,23} (Fig. 3n) and—compared with *Tcf7*^{low} cells—had reduced expression of *Pdcd1* (which encodes PD-1) and *Havcr2* (which encodes TIM3) (Extended Data Fig. 6a). Wild-type *Tcf7*^{high} cells were enriched with the TCF-1 target gene *Slamf6*^{23,24} and memory-like gene signatures^{12,13}, which were further increased in REGNASE-1-null cells (Fig. 3m, Extended Data Fig. 6b). By contrast, wild-type *Tcf7*^{high} cells had a lower expression of *Ifng* and *Gzmb* than did *Tcf7*^{low} cells, but in the absence of REGNASE-1, *Ifng* and *Gzmb* were increased in both *Tcf7*^{high} and *Tcf7*^{low} cells (Extended Data Fig. 6c). Moreover, in REGNASE-1-null OT-I cells, the effector cell factor *Batf*^{25,26}—but not *Id2*²⁷—was highly expressed by both *Tcf7*^{high} and *Tcf7*^{low} cells (Extended Data Fig. 6d, e). Flow cytometry validation revealed that TCF-1 $^+$ cells expressed TOX and SLAMF6 (with modestly

higher levels observed in REGNASE-1-null cells), but had low levels of expression of KLRG1 and TIM3 and intermediate levels of expression of PD-1 (Extended Data Fig. 6f). Collectively, these results establish the dual roles of REGNASE-1 in coordinating T cell effector function and memory-like features in antitumour immunity.

REGNASE-1–BATF shapes effector responses

To identify the mechanisms that underlie REGNASE-1 signalling, we took advantage of the extensive accumulation of tumour-infiltrating REGNASE-1-null cells and performed a secondary in vivo genome-scale CRISPR screening by cotransducing OT-I cells with sgRNA targeting *Regnase-1* and the Brie lentiviral genome-scale sgRNA library²⁸ (Fig. 4a). A total of 331 genes were strongly depleted in the screening, including *Slc7a5*²⁹, *Itk*³⁰, *Prkaa1*³¹, *Mapk1*³² and *Tbx21*³³ (Extended Data Fig. 7a, Supplementary Table 3). Given the role of REGNASE-1 in inhibiting gene expression^{8,9}, we applied two criteria to identify the functional targets of REGNASE-1: candidates should be upregulated in REGNASE-1-null cells in RNA-seq data, but depleted in tumour-infiltrating lymphocytes in the genome-scale CRISPR screening. This analysis revealed two candidates, one of which was *Batf*^{25,26} (Extended Data Fig. 7b). REGNASE-1-null cells showed increased expression of BATF (Fig. 4b, Extended Data Fig. 6d) and enrichment of BATF-binding motifs and gene targets²⁶ (Extended Data Figs. 2g, 7c, d). We next determined whether *Batf* mRNA is regulated by REGNASE-1, using the 3' untranslated region (UTR) of the *Il2* and *Il4* genes as positive and negative controls, respectively⁹ (Extended Data Fig. 7e). The 3' UTR of *Batf* gene was dose-dependently inhibited by REGNASE-1, but not by the nuclease-inactive mutant REGNASE-1(D141N)

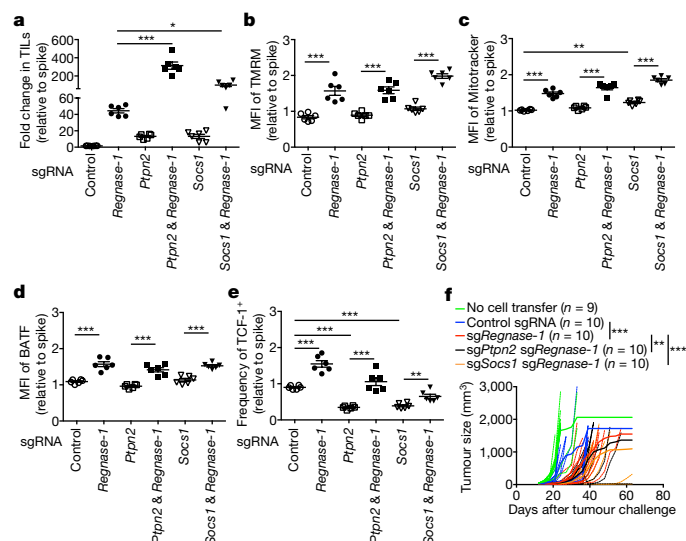


Fig. 5 | Genome-scale CRISPR screening identifies PTPN2 and SOCS1 as additional targets for enhancing the antitumour activity of REGNASE-1-null CD8⁺ T cells. **a–e**, OT-I cells transduced with non-targeting control sgRNA (spike) were mixed at a 1:1 ratio with cells transduced with non-targeting control sgRNA, sgRNA against *Regnase-1*, sgRNA against *Ptpn2*, sgRNAs against *Ptpn2* and *Regnase-1*, sgRNA against *Socs1* or sgRNAs against *Socs1* and *Regnase-1*, and transferred into tumour-bearing hosts individually ($n = 6$). Tumour-infiltrating OT-I cells were analysed at day 7 for quantification of relative OT-I cell percentage in CD8⁺ cells normalized to spike (**a**), quantification of relative MFI of TMRM (**b**), Mitotracker (**c**) and BATF (**d**) normalized to spike, and quantification of the relative frequency of TCF-1⁺ cells normalized to spike (**e**). **f**, Pmel-1 cells transduced with an individual control sgRNA, sgRNA against *Regnase-1* or with two sgRNAs (against *Regnase-1* and either *Ptpn2* (sg*Ptpn2*) or *Socs1* (sg*Socs1*)) (4×10^6 cells in each group) were transferred into mice at day 12 after engraftment of B16 F10 melanoma, followed by analysis of tumour size. Non-treatment control mice received no transfer of T cells. Mean \pm s.e.m. (**a–e**). * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. One-way ANOVA (**a–e**) or two-way ANOVA (**f**). Data in **a–f** are pooled from two independent experiments.

(Fig. 4c), which reveals that BATF is a target of REGNASE-1. Importantly, co-deletion of BATF (Extended Data Fig. 7f, g) markedly reduced the accumulation of REGNASE-1-null OT-I cells in both the periphery and tumour (Fig. 4d, Extended Data Fig. 7h), associated with increased active caspase-3 (Extended Data Fig. 7i). By contrast, cells deficient in both BATF and REGNASE-1 still had increased TCF-1 expression compared to wild-type cells (Extended Data Fig. 7j), which suggests a dispensable role of BATF in TCF-1 expression. Moreover, BATF co-deletion blocked the increased IFN γ production in REGNASE-1-null cells, and dampened the antitumour effects of REGNASE-1-null cells in the pmel-1 ACT model (Extended Data Fig. 7k, l). Therefore, REGNASE-1 targets BATF to impair the accumulation and effector function, but not the TCF-1 expression, of tumour-specific T cells.

BATF expression was aberrantly induced in REGNASE-1-null cells in response to TCR and to, a lesser extent, IL-2—but not IL-21 (Extended Data Fig. 8a). To test whether BATF is a limiting factor in antitumour responses, we transduced wild-type OT-I cells with BATF (Extended Data Fig. 8b) and found that BATF overexpression improved cell accumulation in the spleen (Extended Data Fig. 8c, d) and even more markedly in the tumour (Fig. 4e, Extended Data Fig. 8c). Accordingly, BATF-overexpressing OT-I cells in the tumour had increased cell proliferation and modestly reduced active caspase-3 (Extended Data Fig. 8e, f), and produced more IFN γ , GZMB and TNF (but not IL-2) (Extended Data Fig. 8g). By contrast (but consistent with the role of BATF in promoting effector differentiation²⁶), TCF-1 expression was reduced in cells that

overexpress BATF (Extended Data Fig. 8h). Therefore, BATF is regulated by REGNASE-1 and immune signals and acts as an important rheostat in mediating antitumour effector responses.

To determine the contribution of aberrant BATF expression to the altered chromatin accessibility in REGNASE-1-null cells, we performed ATAC-seq analysis of wild-type, REGNASE-1-null cells, BATF-null cells, and cells deficient for both REGNASE-1 and BATF isolated from tumour-infiltrating lymphocytes. We identified 7,480 genes with increased chromatin accessibility in REGNASE-1-null cells as compared to wild-type cells (Extended Data Fig. 9a), and BATF co-deletion reversed the upregulation of a large proportion of these genes (5,052 genes in total) (Extended Data Fig. 9a). In addition, 2,527 of these 5,052 genes showed downregulated chromatin accessibility in BATF-null cells as compared to wild-type cells (Extended Data Fig. 9a). Thus, a large majority of the genes with increased chromatin accessibility in REGNASE-1-null cells are BATF-dependent.

We next determined the functional pathways by which REGNASE-1 regulates antitumour immunity. Functional enrichment of the top-ranking depleted genes of the genome-scale CRISPR screening revealed the oxidative phosphorylation (OXPHOS) hallmark as the top-enriched pathway (Extended Data Fig. 9b). OXPHOS was also the top-ranking gene set enriched in tumour-infiltrating REGNASE-1-null cells relative to wild-type cells (Extended Data Figs. 4a, 9c). Although mitochondrial metabolism correlates with T cell fitness and antitumour activity^{34,35}, the negative signals involved—especially in the TME—remain unknown. REGNASE-1-null cells showed increased mitochondrial fitness, as indicated by increased mitochondrial mass, membrane potential and volume (Fig. 4f, Extended Data Fig. 9d), as well as higher basal and maximal oxygen consumption rates (Extended Data Fig. 9e). Compared with REGNASE-1-null cells, cells deficient in both BATF and REGNASE-1 downregulated hallmarks associated with OXPHOS and cell cycling (Extended Data Fig. 9f, g). Moreover, BATF co-deletion largely blocked the increased mitochondrial mass and membrane potential in REGNASE-1-null cells at day 5 and day 7 after adoptive transfer (Fig. 4g, Extended Data Fig. 9h). Conversely, BATF overexpression was sufficient to upregulate mitochondrial mass and membrane potential (Fig. 4h). These results collectively reveal a role of BATF in linking REGNASE-1 function and mitochondrial fitness.

To understand the molecular basis for the REGNASE-1- and BATF-mediated regulation of mitochondrial fitness, we mined our ATAC-seq data for altered chromatin accessibility of mitochondrial genes. A total of 341 mitochondrial genes showed significantly upregulated chromatin accessibility in the absence of REGNASE-1, and 214 of these genes were blocked by BATF co-deletion (Extended Data Fig. 9i). Moreover, 96 of these 214 genes showed downregulated chromatin accessibility in BATF-null cells as compared to wild-type cells (Extended Data Fig. 9i). These results further support a crucial contribution of BATF to the enhanced mitochondrial function in the absence of REGNASE-1.

Combination therapy with PTPN2 and SOCS1

Combination therapy is key to the clinical success of cancer immunotherapies³⁶. To identify whether the therapeutic potential of REGNASE-1-null CD8⁺ T cells could be further potentiated, we focused on the top two genes that were enriched in tumour-infiltrating lymphocytes in our genome-scale CRISPR screening: *Ptpn2* and *Socs1* (Extended Data Fig. 7a). We validated the effects of co-deletion of these genes to enhance the accumulation of tumour-infiltrating REGNASE-1-null cells (Fig. 5a, Extended Data Fig. 10a). Deletion of PTPN2 or SOCS1 alone resulted in a modestly increased accumulation of OT-I cells in the tumour (Fig. 5a). Of note, SOCS1 has previously been identified to restrain human T cell proliferation in vitro³⁷, and PTPN2 deletion sensitizes cancer cells to immune checkpoint therapy³⁸. Unlike BATF expression, the expression of PTPN2 or SOCS1 was not affected by

REGNASE-1 deletion (Extended Data Fig. 10b). Deletion of PTPN2 or SOCS1 alone either did not affect, or slightly increased, mitochondrial mass and membrane potential (Fig. 5b, c), and co-deletion of REGNASE-1 still increased these mitochondrial profiles (Fig. 5b, c). Furthermore, deletion of PTPN2 or SOCS1 did not affect BATF expression (Fig. 5d), but significantly reduced TCF-1 expression (Fig. 5e); REGNASE-1 co-deletion was still capable of upregulating TCF-1 expression (Fig. 5e). These comparative analyses reveal that the mechanisms exerted by PTPN2 or SOCS1 are largely discrete from those of REGNASE-1, including the effects on mitochondrial fitness and the regulation of BATF and TCF-1 expression.

We assessed the therapeutic efficacy of pmel-1 T cells with co-deletion of PTPN2 and REGNASE-1 or of SOCS1 and REGNASE-1. Although REGNASE-1 deletion alone greatly improved the therapeutic efficacy, pmel-1 T cells double-deficient for PTPN2 and REGNASE-1 or SOCS1 and REGNASE-1 exhibited additional effects at delaying tumour growth (Fig. 5f). Altogether, our CRISPR screening identifies targets that can be potentially combined with REGNASE-1 deletion in cancer immunotherapy.

Discussion

There is a great need to understand how cell-fate decisions occur in tumour-specific CD8⁺ T cells. Here we reveal that tumour-specific CD8⁺ T cells can be reprogrammed in the TME to acquire extensive accumulation and increased features associated with naive or memory cells for long-term persistence, while still retaining robust effector function (Extended Data Fig. 10c). REGNASE-1 is a major regulator that can be targeted to unleash this unique reprogramming in the TME, with marked therapeutic efficacy against both solid and blood cancers in ACT. The specific transcriptional adaptation of REGNASE-1-null CD8⁺ T cells in the TME highlights a previously unappreciated function of REGNASE-1 after initial T cell activation^{8,9}, to enable the precise temporal and spatial control of T cell responses.

Despite the recent emphasis on the metabolic control of T cell activation and differentiation, the metabolic reprogramming and the molecular events involved remain to be explored in antitumour T cell responses³. Our results highlight the fact that REGNASE-1 restrains mitochondrial metabolism and effector responses through a key gene target, BATF. BATF acts as a limiting factor for programming antitumour responses and mitochondrial metabolism, thereby advancing our understanding of context-dependent roles of the pioneer factor BATF in adaptive immunity^{26,39}. The genome-scale CRISPR screening also reveals PTPN2 and SOCS1 as potential targets to combine with REGNASE-1 deletion to boost antitumour immunity. From a therapeutic perspective, our findings have identified targets for ACT against both solid and blood cancers and point to avenues to reprogramming T cell state and metabolism in cancer immunity and immunotherapy.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1821-z>.

1. Lim, W. A. & June, C. H. The principles of engineering immune cells to treat cancer. *Cell* **168**, 724–740 (2017).
2. Gattinoni, L. et al. Acquisition of full effector function *in vitro* paradoxically impairs the *in vivo* antitumor efficacy of adoptively transferred CD8⁺ T cells. *J. Clin. Invest.* **115**, 1616–1626 (2005).
3. Kishton, R. J., Sukumar, M. & Restifo, N. P. Metabolic regulation of T cell longevity and function in tumor immunotherapy. *Cell Metab.* **26**, 94–109 (2017).

4. Muri, J. et al. The thioredoxin-1 system is essential for fueling DNA synthesis during T-cell metabolic reprogramming and proliferation. *Nat. Commun.* **9**, 1851 (2018).
5. Peng, M. et al. Aerobic glycolysis promotes T helper 1 cell differentiation through an epigenetic mechanism. *Science* **354**, 481–484 (2016).
6. Vanoica, L. et al. Conditional deletion of ferritin H in mice reduces B and T lymphocyte populations. *PLoS ONE* **9**, e89270 (2014).
7. Quyang, W., Beckett, O., Flavell, R. A. & Li, M. O. An essential role of the Forkhead-box transcription factor Foxo1 in control of T cell homeostasis and tolerance. *Immunity* **30**, 358–371 (2009).
8. Matsushita, K. et al. Zc3h12a is an RNase essential for controlling immune responses by regulating mRNA decay. *Nature* **458**, 1185–1190 (2009).
9. Uehata, T. et al. Malt1-induced cleavage of regnase-1 in CD4⁺ helper T cells regulates immune activation. *Cell* **153**, 1036–1049 (2013).
10. Churchman, M. L. et al. Synergism of FAK and tyrosine kinase inhibition in Ph⁺ B-ALL. *JCI Insight* **1**, 86082 (2016).
11. Singer, M. et al. A distinct gene module for dysfunction uncoupled from activation in tumor-infiltrating T cells. *Cell* **166**, 1500–1511 (2016).
12. Im, S. J. et al. Defining CD8⁺ T cells that provide the proliferative burst after PD-1 therapy. *Nature* **537**, 417–421 (2016).
13. Leong, Y. A. et al. CXCR5⁺ follicular cytotoxic T cells control viral infection in B cell follicles. *Nat. Immunol.* **17**, 1187–1196 (2016).
14. Zhou, X. et al. Differentiation and persistence of memory CD8⁺ T cells depend on T cell factor 1. *Immunity* **33**, 229–240 (2010).
15. Hurlton, L. V. et al. Tethered IL-15 augments antitumor activity and promotes a stem-cell memory subset in tumor-specific T cells. *Proc. Natl Acad. Sci. USA* **113**, E7788–E7797 (2016).
16. Roychoudhuri, R. et al. BACH2 regulates CD8⁺ T cell differentiation by controlling access of AP-1 factors to enhancers. *Nat. Immunol.* **17**, 851–860 (2016).
17. Ichii, H., Sakamoto, A., Kuroda, Y. & Tokuhisa, T. Bcl6 acts as an amplifier for the generation and proliferative capacity of central memory CD8⁺ T cells. *J. Immunol.* **173**, 883–891 (2004).
18. Man, K. et al. Transcription factor IRF4 promotes CD8⁺ T cell exhaustion and limits the development of memory-like T cells during chronic infection. *Immunity* **47**, 1129–1141 (2017).
19. Sade-Feldman, M. et al. Defining T cell states associated with response to checkpoint immunotherapy in melanoma. *Cell* **175**, 998–1013 (2018).
20. Buenostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
21. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
22. Khan, O. et al. TOX transcriptionally and epigenetically programs CD8⁺ T cell exhaustion. *Nature* **571**, 211–218 (2019).
23. Miller, B. C. et al. Subsets of exhausted CD8⁺ T cells differentially mediate tumor control and respond to checkpoint blockade. *Nat. Immunol.* **20**, 326–336 (2019).
24. Utzschneider, D. T. et al. T cell factor 1-expressing memory-like CD8⁺ T cells sustain the immune response to chronic viral infections. *Immunity* **45**, 415–427 (2016).
25. Ciofani, M. et al. A validated regulatory network for Th17 cell specification. *Cell* **151**, 289–303 (2012).
26. Kurachi, M. et al. The transcription factor BATF operates as an essential differentiation checkpoint in early effector CD8⁺ T cells. *Nat. Immunol.* **15**, 373–383 (2014).
27. Yang, C. Y. et al. The transcriptional regulators Id2 and Id3 control the formation of distinct memory CD8⁺ T cell subsets. *Nat. Immunol.* **12**, 1221–1229 (2011).
28. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
29. Sinclair, L. V. et al. Control of amino-acid transport by antigen receptors coordinates the metabolic reprogramming essential for T cell differentiation. *Nat. Immunol.* **14**, 500–508 (2013).
30. Atherly, L. O., Brehm, M. A., Welsh, R. M. & Berg, L. J. Tec kinases Itk and Rlk are required for CD8⁺ T cell responses to virus infection independent of their role in CD4⁺ T cell help. *J. Immunol.* **176**, 1571–1581 (2006).
31. Blagih, J. et al. The energy sensor AMPK regulates T cell metabolic adaptation and effector responses *in vivo*. *Immunity* **42**, 41–54 (2015).
32. D'Souza, W. N., Chang, C. F., Fischer, A. M., Li, M. & Hedrick, S. M. The Erk2 MAPK regulates CD8 T cell proliferation and survival. *J. Immunol.* **181**, 7617–7629 (2008).
33. Sullivan, B. M., Juedes, A., Szabo, S. J., von Herrath, M. & Glimcher, L. H. Antigen-driven effector CD8 T cell function regulated by T-bet. *Proc. Natl Acad. Sci. USA* **100**, 15818–15823 (2003).
34. Geiger, R. et al. L-Arginine modulates T cell metabolism and enhances survival and antitumor activity. *Cell* **167**, 829–842 (2016).
35. Kawalekar, O. U. et al. Distinct signaling of coreceptors regulates specific metabolism pathways and impacts memory development in CAR T cells. *Immunity* **44**, 380–390 (2016).
36. Sharma, P. & Allison, J. P. The future of immune checkpoint therapy. *Science* **348**, 56–61 (2015).
37. Shifrut, E. et al. Genome-wide CRISPR screens in primary human T cells reveal key regulators of immune function. *Cell* **175**, 1958–1971 (2018).
38. Manguso, R. T. et al. *In vivo* CRISPR screening identifies *Ptpn2* as a cancer immunotherapy target. *Nature* **547**, 413–418 (2017).
39. Quigley, M. et al. Transcriptional analysis of HIV-specific CD8⁺ T cells shows that PD-1 inhibits T cell function by upregulating BATF. *Nat. Med.* **16**, 1147–1151 (2010).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Cell lines and mice

The B16 F10 cell line was purchased from ATCC. The B16 Ova cell line was provided by D. Vignali. The human CD19-Ph⁺ B-ALL cell line was provided by T. Geiger (manuscript in preparation). C57BL/6, OT-I, pmel-1 and *Rosa26-Cas9* knock-in mice⁴⁰ were purchased from The Jackson Laboratory. CAR T transgenic mice (T cells express CARs that consist of anti-human CD19 (human CD19) scFv fragments, CD8 transmembrane domain and 4-1BB-CD3 ζ signalling tail) were provided by T. Geiger (manuscript in preparation). We crossed *Rosa26-Cas9* knock-in mice⁴⁰ with OT-I⁴¹, pmel-1⁴² or CAR T transgenic mice to express Cas9 in antigen-specific CD8⁺ T cells. Sex-matched mice were used at 7–16 weeks old unless otherwise noted. All mice were kept in a specific-pathogen-free facility in the Animal Resource Center at St Jude Children's Research Hospital. Experiments and procedures were performed in accordance with the Institutional Animal Care and Use Committee of St Jude Children's Research Hospital.

Cell purification and viral transduction

Naive Cas9-expressing OT-I cells were isolated from the spleen and peripheral lymph nodes (PLNs) of Cas9-OT-I mice using a naive CD8 α^+ T cell isolation kit (Miltenyi Biotec 130-096-543) according to the manufacturer's instructions. Purified naive OT-I cells were activated in vitro for 18 h with 10 μ g/ml anti-CD3 (2C11; Bio X Cell), 5 μ g/ml anti-CD28 (37.51; Bio X Cell) before viral transduction. Viral transduction was performed by spin-infection at 800g at 25 °C for 3 h with 10 μ g/ml polybrene (Sigma). Cells were cultured with human IL-2 (20 IU/ml; PeproTech), mouse IL-7 (2.5 ng/ml; PeproTech) and IL-15 (25 ng/ml; PeproTech) for 3–4 days. Transduced cells were sorted using a Reflection cell sorter (iCyt) before adoptive transfer into recipients. sgRNAs were designed by using the online tool (<https://portals.broadinstitute.org/gpp/public/analysis-tools/sgRNA-design>). sgRNAs used in this study were as follows: non-targeting control sgRNA, ATGACACTTACGGTACTCGT; *Regnase-1* sgRNA, AAGGCAGTGGTTCTTACGA; *Regnase-1* sgRNA no. 2, GGAGTG-GAAACGCTTCATCG; *Batf* sgRNA, AGAGATCAAACAGCTACCG; *Batf* sgRNA no. 2, AGGACTCATCTGATGATGTG (which gave results similar to those of *Batf* sgRNA; data not shown); *Ptpn2* sgRNA, AAGAAGTTCATCTTAACAC; *Ptpn2* sgRNA no. 2: CACTCTATGAGGATAGTCAT (which gave results similar to those of *Ptpn2* sgRNA; data not shown); *Socs1* sgRNA, TGATGCGCCGGTAAATCGGAG; *Socs1* sgRNA no. 2, TGGTGGCG-GACAGTCGCCAA (which gave results similar to those of *Socs1* sgRNA; data not shown). The coding sequence of *Batf* (Addgene no. 34575) was subcloned into pMIG-II retroviral vector (Addgene no. 52107), which was co-transfected into Plat-E cells with the helper plasmid pCL-Eco (Addgene no. 12371) for the production of retrovirus.

Lentiviral sgRNA metabolic library CRISPR–Cas9 mutagenesis screening

Lentiviral and retroviral sgRNA vector design. The lentiviral sgRNA vector was generated from lentiGuide-puro vector by replacing the EF-1 α PuroR fragment with a mouse PGK promoter-driven ametrine (or GFP or mCherry) fluorescent protein. The retroviral sgRNA vector was generated from pLMPd-Amt vector⁴³ by replacing the miR30 shRNA cassette with the U6-promoter-driven gRNA cassette from the lentiGuide-puro vector.

Lentiviral sgRNA metabolic library construction. The gene list (3,017 genes) of the mouse metabolic library was based on reported human metabolic genes⁴⁴. A total of six gRNAs were designed for each mouse metabolic gene according to previously published selection criteria⁴⁵

and were split into two sub-libraries (AAAQ05 and AAAR07; Supplementary Table 1), each containing 500 non-targeting controls. Oligonucleotides containing the guide sequence were synthesized (Custom Array), PCR-amplified and cloned into the recipient vector via a Golden Gate cloning procedure, including 5 μ l Tango Buffer (ThermoFisher), 5 μ l dithiothreitol (10 mM stock), 5 μ l ATP (10 mM stock), 500 ng vector (pre-digested with Esp3I, gel-extracted and isopropanol-precipitation-purified), 100 ng insert PCR product, 1 μ l Esp3I (ThermoFisher ER0452), 1 μ l T7 ligase (Enzymatics, 3,000 U/ μ l, L6020L) and water, up to 50 μ l, and incubated in a cycle (5 min at 37 °C and 5 min at 20 °C) for 100 times. The product was then purified by isopropanol precipitation and electroporated into STBL4 cells (Life Technologies 11635018). The distribution of the library was determined by Illumina sequencing.

In vivo screening. Lentivirus was produced by co-transfecting HEK293T cells with the lentiviral metabolic library plasmids, psPAX2 (Addgene plasmid no. 12260) and pCAG4-Eco. At 48 h after transfection, virus was collected and frozen at –80 °C. Four hundred to five hundred million naive Cas9-expressing OT-I cells were isolated from 8–14 Cas9 OT-I mice and transduced at a multiplicity of infection of 0.3 to achieve about 20% transduction efficiency. After viral transduction, cells were cultured with human IL-2 (20 IU/ml; PeproTech), mouse IL-7 (2.5 ng/ml; PeproTech) and IL-15 (25 ng/ml; PeproTech) for 4 days to allow gene editing to occur. Transduced cells expressing ametrine were sorted using a Reflection cell sorter (iCyt), and an aliquot of 5×10^6 transduced OT-I cells was saved as input (about 500 \times cell coverage per sgRNA). Transduced OT-I cells (5×10^6 cells per recipient) were intravenously transferred into mice at day 14 after B16 Ova melanoma engraftment. Sixty recipients were randomly divided into three groups as biological replicates in each sub-library screening. At 7 days after adoptive transfer, transferred ametrine⁺ OT-I cells were recovered from the tumour pooled from 20 recipients per sample using a Reflection cell sorter (iCyt). On average, 5×10^5 OT-I cells per sample (about 50 \times cell coverage per sgRNA) were recovered and used for deep sequencing of the sgRNA cassette, with the expectation that sgRNAs capable of improving ACT should be enriched in tumour-infiltrating OT-I cells.

Sequencing library preparation. Genomic DNA was extracted by using the DNeasy Blood & Tissue Kits (Qiagen 69506). Primary PCR was performed by using the KOD Hot Start DNA Polymerase (Millipore 71086) and the following pair of Nextera next-generation sequencing (NGS) primers (Nextera NGS forward (-F): TCGTCGGCAGCGTCAGATGTGATAAGAGACAGTgttgaaaggacgaacacccg; Nextera NGS reverse (-R): GTCTCGTGGGCTCGGAGATGTGATAAGAGACAGccacttttcaagtgataacgg). Primary PCR products were purified using the AMPure XP beads (Beckman A63881). A second PCR was performed to add adaptors and indexes to each sample. Hi-seq 50-bp single-end sequencing (Illumina) was performed.

Data processing. For data analysis, FASTQ files obtained after sequencing were demultiplexed using the HiSeq Analysis software (Illumina). Single-end reads were trimmed and quality-filtered using the CLC Genomics Workbench v.11 (Qiagen) and matched against sgRNA sequences from the sgRNA metabolic library. Read counts for sgRNAs were normalized against total read counts across all samples. For each sgRNA, the fold change (\log_2 -transformed ratio) for enrichment was calculated between each of the biological replicates and the input experiment. After merging the quantification results from two sub-libraries, candidate genes were ranked on the basis of the average enrichment of their six gene-specific sgRNAs in tumour-infiltrating OT-I cells relative to input (\log_2 (TIL/input ratio); adjusted $P < 0.05$). The gene-level false-discovery-rate-adjusted P value was calculated among multiple sgRNAs ($n = 6$) of each gene, using a two-tailed paired Student's t -test between \log_2 -transformed average normalized read counts of tumour

Article

samples and those of input sample, and the *P* value was further adjusted using Bonferroni correction with gene size.

Genome-scale sgRNA Brie library CRISPR–Cas9 mutagenesis screening

In vivo screening. Lentivirus was produced by co-transfecting HEK293T cells with lentiviral genome-scale Brie library plasmids with the puromycin-resistance gene²⁸, psPAX2 and pCAG4-Eco. At 48 h after transfection, virus was collected and frozen at -80°C . Two hundred million Cas9-expressing OT-I cells were isolated from 12 Cas9 OT-I mice and co-transduced with the Brie sgRNA library and *Regnase-1* sgRNA–ametrine. After viral transduction, cells were cultured with human IL-2 (20 IU/ml; PeproTech), mouse IL-7 (2.5 ng/ml; PeproTech) and IL-15 (25 ng/ml; PeproTech) for 2 days. Brie-sgRNA-library-transduced cells were then selected by culture with 4 $\mu\text{g}/\text{ml}$ puromycin in the presence of the abovementioned cytokines for another 3 days. Following puromycin selection, ametrine⁺ cells were sorted using a Reflection cell sorter (iCyt) to select for cells cotransduced with *sgRegnase-1* and Brie-library sgRNAs, and an aliquot of 10×10^6 transduced OT-I cells was saved as input (about 120 \times cell coverage per sgRNA). The majority of the co-transduced OT-I cells (5×10^6 cells per recipient) were then intravenously transferred into mice at day 14 after B16 Ova melanoma engraftment. Twenty recipients were randomly divided into two groups as biological replicates. At 7 days after adoptive transfer, transferred ametrine⁺ OT-I cells were recovered from the tumour pooled from 10 recipients per sample using a Reflection cell sorter (iCyt). On average, 3×10^6 OT-I cells per sample (about 40 \times cell coverage per sgRNA) were recovered. DNA extraction and sequencing library preparation were as described in ‘Sequencing library preparation’.

Data processing. For data analysis, FASTQ files obtained after sequencing were demultiplexed using the HiSeq Analysis software (Illumina). *Regnase-1* sgRNA (GGAGTGGAAACGCTTCATCG) reads were removed, and single-end reads were trimmed and quality-filtered using the CLC Genomics Workbench v.11 (Qiagen) and matched against sgRNA sequences from the genome-scale sgRNA Brie library. Read counts for sgRNAs were normalized against total read counts across all samples. For each sgRNA, the fold change (\log_2 -transformed ratio) for enrichment was calculated between each of the biological replicates and the input experiment. Gene ranking was based on the average enrichment ($\log_2(\text{TIL}/\text{input ratio})$) among replicates in representation of four individual corresponding sgRNAs in the genome-scale sgRNA Brie library. The gene-level false-discovery-rate-adjusted *P* value was calculated among multiple sgRNAs ($n = 4$) of each gene, using a two-tailed paired Student's *t*-test between \log_2 -transformed average normalized read counts of tumour samples and those of the input sample, and the *P* value was further adjusted using Bonferroni correction with gene size.

Flow cytometry

For analysis of surface markers, cells were stained in PBS (Gibco) containing 2% (w/v) BSA (Sigma). Surface proteins were stained for 30 min on ice. Intracellular staining was performed with Foxp3/transcription factor staining buffer set, according to the manufacturer's instructions (eBioscience). Intracellular staining for cytokines was performed with a fixation/permeabilization kit (BD Biosciences). Active caspase-3 staining was performed using instructions and reagents from the Active Caspase-3 Apoptosis Kit (BD Biosciences). BrdU staining (pulsed for 18 h) was performed using instructions and reagents from the APC BrdU Flow Kit (BD Biosciences). 7-AAD (Sigma) or fixable viability dye (eBioscience) was used for dead-cell exclusion. The following antibodies were used: anti-IFN γ (XMGL2), anti-TNF (MAb11), anti-IL-2 (JES6-5H4), anti-CD69 (H1.2F3), anti-CD25 (PC61.5), anti-KLRG1 (2F1), anti-ICOS (7E.17G9), anti-LAG3 (C9B7W), anti-PD-1 (J43), anti-CTLA4 (1B8), anti-TOX (TXRX10), anti-TIM3 (RMT3-23) (all from eBioscience); anti-GZMB (QA16A02), anti-CD49a (HM α 1), anti-CD44 (IM7), anti-Ki-67

(16A8), anti-CD127 (A7R34) (all from Biolegend); anti-BrdU (3D4), anti-active caspase-3 (C92-605), anti-pH2A.X-S139 (N1-431) (DNA damage biomarker, which measures phosphorylation of the histone variant H2A.X at Ser139^{46,47}), anti-SLAMF6 (13G3) (all from BD Biosciences); anti-BATF (D7C5), anti-TCF-1 (C63D9) (all from Cell Signaling Technology); anti-CD8 α (53-6.7) (from SONY); and anti-CD62L (MEL-14) (from TONBO Bioscience). To monitor cell division, lymphocytes were labelled with CellTrace Violet (Life Technologies). For mitochondrial staining, lymphocytes were incubated for 30 min at 37°C with 10 nM Mito Tracker Deep Red (Life Technologies) or 20 nM TMRM (ImmunoChemistry Technologies) after staining surface markers. Flow cytometry data were analysed using Flowjo 9.9.4 (Tree Star).

Adoptive T cell transfer for tumour therapy

B16 Ova cells (2×10^5) or B16 F10 cells (2×10^5) were injected subcutaneously into female C57BL/6 mice (7–10 weeks of age). At day 12, mice bearing tumours of a similar size were randomly divided into 3 groups (5–8 mice per group), and sgRNA-transduced OT-I cells (5×10^6) (for the treatment of B16 Ova melanomas) or pmel-1 (5×10^6) (for the treatment of B16 F10 melanomas) were injected intravenously. Tumours were measured every three days with digital callipers and tumour volumes were calculated by the formula: $\text{length} \times \text{width} \times [(\text{length} \times \text{width})^{0.5}] \times \pi/6$ (ref. ⁴⁸). Death was defined as the point at which a progressively growing tumour reached 15 mm in the longest dimension. For the treatment of human CD19-Ph⁺ B-ALL, mice engrafted with human CD19-Ph⁺ B-ALL cells (1×10^6) were treated at day 7 with sgRNA-transduced CD8⁺ CAR T cells (5×10^6). Mice were imaged using the Xenogen imaging system (Caliper Life Science).

TIL isolation

To isolate TILs, B16 Ova melanoma was excised, minced and digested with 0.5 mg/ml collagenase IV (Roche) + 200 IU/ml DNase I (Sigma) for 1 h at 37°C , and then passed through 70- μm filters to remove undigested tumour tissues. TILs were then isolated by density-gradient centrifugation over Percoll (Life Technologies).

Gene expression profiling and GSEA

OT-I cells transduced with control sgRNA ($n = 4$ biological replicates) and *Regnase-1* sgRNA ($n = 5$ biological replicates) were isolated from the tumours or PLN of the hosts of the in vivo dual colour transfer assay, and analysed with RNA-seq. For RNA-seq, RNA was quantified using the Quant-iT RiboGreen assay (Life Technologies) and quality-checked by 2100 Bioanalyzer RNA 6000 Nano assay (Agilent) or LabChip RNA Pico Sensitivity assay (PerkinElmer) before library generation. Libraries were prepared from total RNA with the TruSeq Stranded Total RNA Library Prep Kit according to the manufacturer's instructions (Illumina, PN 20020595). Libraries were analysed for insert size distribution on a 2100 BioAnalyzer High Sensitivity kit (Agilent Technologies) or Caliper LabChip GX DNA High Sensitivity Reagent Kit (PerkinElmer). Libraries were quantified using the Quant-iT PicoGreen dsDNA assay (Life Technologies) or low-pass-sequencing with a MiSeq nano kit (Illumina). Paired-end 100-cycle sequencing was run on the HiSeq 4000 (Illumina). The raw reads were trimmed for adaptor sequences using Trimmomatic v.0.36 using parameters ILLUMINACLIP:adaptor.fa:2:30:10 LEADING:10 TRAILING:10 SLIDINGWINDOW:4:18 MINLEN:32, followed by mapping to the mm9 reference genome downloaded from gencode release M1 (<https://www.gencodegenes.org/mouse/releases.html>) using star v.2.5.2b. with default parameters. Reads were summarized at gene level using the Python script htseq-count. Differential expression analysis was performed using the R package DESeq2 v.1.18.1. OT-I cells transduced with *Regnase-1* sgRNA ($n = 3$ biological replicates) or *Batf* and *Regnase-1* sgRNAs ($n = 3$ biological replicates) were isolated from the tumours in the in vivo dual colour transfer assay and used for microarray analysis (Affymetrix Mouse Clariom S Assay). For microarray analysis, the expression signals were summarized using the

robust multi-array average algorithm Affymetrix Expression Console v.1.1, followed by differential expression analysis performed using the R package limma v.3.34.9. All the plots were generated using the R package ggplot2 v.2.2.1. Differentially expressed transcripts were identified using lmFit method implemented in limma v.3.34.9 and the Benjamini–Hochberg method was used to estimate the false discovery rate (FDR) as previously described⁴⁹. Differentially expressed genes were defined by $|\text{fold change} (\log_2\text{-transformed ratio})| > 0.5$; Benjamini–Hochberg adjusted $P < 0.05$. GSEA was performed as previously described⁵⁰ using the ‘Hallmark’ database. For GSEA using manually curated gene signatures from public datasets, the microarray dataset (GSE84105)⁵² was used for generating ‘CXCR5’ exhausted CD8 (Ahmed)’ gene signatures ($< 5\%$ FDR); because the total number of upregulated and downregulated genes was more than 200, we ranked genes by their fold change (\log_2 -transformation of their expression in CXCR5⁺ versus CXCR5⁻) and used the top-200 upregulated genes as ‘CXCR5⁺ exhausted CD8 (Ahmed)’. RNA-seq data (GSE76279)⁵³ were processed using DESeq2 R package v.1.16.1 to generate ‘CXCR5⁺ exhausted CD8 (Yu)’ using the same strategy.

ATAC-seq and data analysis

Library preparation. To prepare the ATAC-seq library, tumour-infiltrating sgRNA-transduced OT-I cells were collected in the following two batches: (a) control-sgRNA- and *Regnase-1*-sgRNA-transduced OT-I cells ($n = 4$ biological replicates per group) were isolated from tumour-bearing mice using the *in vivo* dual colour transfer assay; and (b) TIL OT-I cells transduced with control sgRNA, *Regnase-1* sgRNA, *Batf* sgRNA or *Batf* and *Regnase-1* sgRNAs ($n = 2$ –4 replicates per group) were isolated from the tumour-bearing mice that received the individual transfer of sgRNA-transduced OT-I cells. Sorted T cells were incubated in 50 μl ATAC-seq lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl₂, 0.1% IGEPAL CA-630) on ice for 10 min. The resulting nuclei were pelleted at 500g for 10 min at 4 °C. The supernatant was carefully removed with a pipette and discarded. The pellet was resuspended in 50 μl transposase reaction mix (25 μl 2 \times TD buffer, 22.5 μl nuclease-free water and 2.5 μl transposase) and incubated for 30 min at 37 °C. After the reaction, the DNA was cleaned up using the Qiagen MinElute kit. The barcoding reaction was run using the NEBNext HiFi kit based on manufacturer’s instructions, and amplified for 5 cycles according to a previous publication²⁰ and using the same primers. Ideal cycle numbers were determined from 5 μl (of 50 μl) from the previous reaction mix using KAPA SYBRFast (Kapa Biosystems) and a 20-cycle amplification on an Applied Biosystems 7900HT. Optimal cycles were determined from the linear part of the amplification curve and the remaining 45 μl of PCR reaction was amplified in the same reaction mix using the optimal cycle number.

Data analysis. ATAC-seq analysis was performed as previously described⁵¹. In brief, 2 \times 100-bp paired-end reads obtained from all samples were trimming for Nextera adaptor by cutadapt (version 1.9, paired-end mode, default parameter with ‘-m 6 -O 20’) and aligned to mouse genome mm9 downloaded from gencode release M1 (<https://www.encodegenes.org/mouse/releases.html>) by BWA (version 0.7.16, default parameters)⁵², duplicated reads were then marked with Picard (version 2.9.4) and only non-duplicated proper paired reads have been kept by samtools (parameter ‘-q 1 -F 1804’ version 1.9)⁵³. After adjustment of TnS shift (reads were offset were offset by +4 bp for the sense strand and –5 bp for the antisense strand), we separated reads into nucleosome-free, mononucleosome, dinucleosome and trinucleosome (as previously described²⁰) by fragment size and generated .bigwig files by using the centre 80-bp of fragments and scaled to 30 \times 10⁶ nucleosome-free reads. We observed reasonable nucleosome-free peaks and a pattern of mono-, di- and tri-nucleosomes on IGV (version 2.4.13)⁵⁴ and all 8 samples had about 10 \times 10⁶ nucleosome-free reads; we therefore concluded that the data qualities were good. Next, we merged each of the two replicates

to enhance peak-calling on nucleosome-free reads by MACS2 (version 2.1.1.20160309 default parameters with ‘-extsize 200–nomodel’)⁵⁵. To assure replicability, we first finalized nucleosome-free regions for each genotype and retained a peak only if it called with a higher cutoff (macs2 -q 0.05) in one merged sample and at least called with a lower cutoff (macs2 -q 0.5) in the other merged sample. The reproducible peaks were further merged between wild-type and REGNASE-1-null samples and then we counted nucleosome-free reads from each of the eight samples using bedtools (v.2.24.0)⁵⁶. To find the differentially accessible regions, we first normalized raw nucleosome-free read counts using the trimmed mean of M -values normalization method and applied an empirical Bayes statistics test after linear fitting using the voom package (R 3.23, edgeR 3.12.1, limma 3.26.9)⁵⁷. FDR-corrected P value < 0.05 and fold change (\log_2 -transformed ratio) > 0.5 were used as cutoffs for more-accessible or less-accessible regions in REGNASE-1-null samples. We annotated the differentially accessible regions in ATAC-seq data for the nearest genes, and also superimposed these genes with 1,158 mitochondrial genes defined in the MitoCarta 2.0 database⁵⁸. For motif analysis, we further selected regions < 0.05 fold change and P value > 0.5 as control regions. FIMO from MEME suite (version 4.11.3, ‘-thresh 1e-4–motif-pseudo 0.0001’)⁵⁹ was used for scanning motif (TRANSFAC database, only included Vertebrata and not 3D structure-based) matches in the nucleosome-free regions and two-tailed Fisher’s exact test was used to test whether a motif is significant enriched for differentially accessible regions compared to the control regions.

Footprinting of transcription-factor binding sites. Footprinting was performed as previously described⁵¹. In brief, we first generated .bigwig files according to all tags of adjusted reads, and then normalized them according to the number of autosome reads to 2 \times 10⁸ reads (for example, a sample with 1 \times 10⁸ autosome reads would be scaled so as to double the bigwig profile). We then generated average .bigwig files from the mean of replicates at each base pair for each sample, using motif matches within a nucleosome-free region for footprinting and taking the average profile across all motif matches at each base pair from –100 bp from motif match centres to +100 bp. Finally, the footprinting profiles were smoothed with 10-bp bins and plotted using deeptools (v.2.5.7)⁶⁰.

To identify the enrichment of BATF binding motifs, nucleosome-free differentially accessible regions were defined at $|\text{fold change} (\log_2\text{-transformed ratio})| > 0.5$; $P < 0.05$, and the peaks were further annotated as more- or less-accessible regions in REGNASE-1-null OT-I cells compared to wild-type controls. For each group, differentially accessible peaks were overlapped with BATF chromatin immunoprecipitation with sequencing (ChIP-seq) peaks (downloaded from GSE54191²⁶) to identify the common regions between ATAC-seq peaks and BATF ChIP-seq peaks using bedtools (version 2.25.0). Finally, FIMO⁶¹ from MEME suite (version 4.9.0) was used to scan the overlapping regions with TRANSFAC motifs associated with BATF to identify the number of motifs enriched in the differentially accessible regions in REGNASE-1-null (shown as ‘# Match (REGNASE-1-null)’ in Extended Data Fig. 7d) or wild-type control samples (shown as ‘# Match (wild-type)’ in Extended Data Fig. 7d), and Fisher’s exact test was used to test the significance of enrichment. This statistical bioinformatics method has successfully been used to circumvent cell-number limitations^{51,62}.

Imaging

B16 Ova melanomas were fixed in PBS containing 2% PFA, 0.3% Triton-100 and 1% DMSO for 24 h before cryoprotection in 30% sucrose. Cryosections were blocked with 1% BSA and 0.05% Tween-20 in TBS (20 mM Tris, pH 8.0 and 100 mM NaCl) for 1 h at room temperature before overnight incubation in blocking buffer containing the following antibodies; anti-mCherry (Biorbyt orb11618), anti-GFP (Rockland Immuno 600-401-215), anti-TCF-7 (C63D9) (Cell Signaling Technology 2203)

Article

and anti-TOM20 (2F8.1) (Millipore MABT166). Slides were washed in TBS before application of AF488, Cy3 or AF647 secondary antibodies (Jackson Immuno) for 1 h at room temperature before mounting with Prolong Diamond hardset medium containing DAPI (Thermo Fisher). Widefield fluorescence microscopy was performed using a motorized Nikon TiE inverted microscope equipped with a 20× Plan Apo 0.75 NA objective, standard DAPI, FITC and TRITC filter sets and an EMCCD camera (Andor). The entire tissue section was stitched on the basis of the DAPI fluorescent signal and the subsequent large images were analysed using NIS Elements software (Nikon Instruments). Images were segmented per channel, and further refined using a spot identification algorithm to identify single cells and positional information within the tumour. The number of cells per square area was determined following manual delineation of the tumour border. Analysis of transcription factor localization was performed using a Marianis spinning disk confocal microscope (Intelligent Imaging Innovations) equipped with a 100× 1.4 NA objective and Prime 95B sCMOS camera, and analysed using Slidebook software (Intelligent Imaging Innovations).

RNA isolation and real-time PCR

RNA was isolated using the RNeasy Micro Kit (Qiagen 74004) following the manufacturer's instructions. RNA was converted to cDNA using the High Capacity cDNA Reverse Transcription Kit (ThermoFisher 4368813) according to manufacturer's instructions. Real-time PCR was performed on the QuantStudio 7 Flex System (Applied Biosystems) using the PowerSYBR Green PCR Master Mix (ThermoFisher 4367659) and the following primers: *Irf4*-F: TCCGACAGTGGTTGATCGAC, *Irf4*-R: CCTCAGATTGTAGTCTGCTT.

Protein extraction and immunoblot

Cells were lysed in RIPA buffer (ThermoFisher 89900), resolved in 4–12% Criterion XT Bis-Tris Protein Gel (Bio-Rad 3450124) and transferred to PVDF membrane (Bio-Rad 1620177). Membranes were blocked using 5% BSA for 1 h and then incubated for overnight with anti-MCPIP1 (604421) (R&D), anti-BATF (D7C5) (Cell Signaling Technology), anti-PTPN2 (E-11) (Santa Cruz Biotechnology), anti-SOCS1 (E-9) (Santa Cruz Biotechnology), anti-HSP90 (MAB3286) (R&D) or anti- β -actin (8H10D10) (Cell Signaling Technology) antibody. Membranes were washed 6 times with TBST and then incubated with 1:5,000 diluted HRP-conjugated anti-mouse IgG (W4021) (from Promega) for 1 h. Following another 6 washes with TBST, the membranes were imaged using the ODYSSEY Fc Analyzer (LI-COR).

Luciferase assay

The full-length 3' UTR constructs of *Batf* (MmiT031430-MT06), *Il2* (MmiT092987-MT06) and *Il4* (MmiT092992-MT06) mRNAs were purchased from GeneCopoeia, each containing two luciferase genes: firefly luciferase gene for the 3' UTR of the targeted gene, and the Renilla luciferase gene as an internal control. The cDNA of wild-type REGNASE-1 (Dharmacon MMM1013-202800061) was cloned into the pMIG-II vector. The D141N mutant of REGNASE-1 was generated by site-directed mutagenesis using the KOD Hot Start DNA Polymerase (Millipore 71086). HEK293T cells were transfected with the 3' UTR construct of interest together with wild-type or the D141N mutant of REGNASE-1 expression plasmid, or empty control plasmid. At 48 h after transfection, cells were lysed and luciferase activities in the lysates were determined using the Luc-Pair Duo-Luciferase Assay Kit (GeneCopoeia LF002) according to manufacturer's instructions.

Seahorse metabolic assay

Oxygen consumption rates were measured in XF medium under basal conditions and in response to 1 μ M oligomycin, 1.5 μ M fluoro-carbonyl cyanide phenylhydrazone (FCCP) and 500 nM rotenone using an XF96 Extracellular Flux Analyzer (EFA) (Seahorse Bioscience).

scRNA-seq and data analysis

Library preparation. Control-sgRNA- and *Regnase-1*-sgRNA-transduced OT-I cells were sorted on a Reflection cell sorter (iCyt) from TILs pooled from the in vivo dual transfer hosts (6–8 mice per sample) at day 7 after adoptive transfer into tumour-bearing mice. The cells were counted and examined for viability using a Luna Dual Fluorescence Cell Counter (Logos Biosystems). All samples were spun down at 2,000 rpm for 5 min. The supernatant was removed, and cells were re-suspended in 100 μ l of 1× PBS (Thermo Fisher Scientific) + 0.04% BSA (Amresco). The cells were then counted and examined for viability using a Luna Dual Fluorescence Cell Counter (Logos Biosystems). Cell counts were about 1×10^6 cells per millilitre and viability was above 98%. Single-cell suspensions were loaded onto the Chromium Controller according to their respective cell counts to generate 6,000 single-cell gel beads in emulsion per sample. Each sample was loaded into a separate channel. Libraries were prepared using the Chromium Single Cell 3' v2 Library and Gel Bead Kit (10X Genomics). The cDNA content of each sample after cDNA amplification of 12 cycles was quantified and quality-checked using a High-Sensitivity DNA chip with a 2100 Bioanalyzer (Agilent Technologies) to determine the number of PCR amplification cycles to yield a sufficient library for sequencing. After library quantification and quality-checking using DNA 1000 chip (Agilent Technologies), samples were diluted to 3.5 nM for loading onto the HiSeq 4000 (Illumina) with a 2 × 75-bp paired-end kit using the following read length: 26-bp read 1, 8-bp i7 index, and 98-bp read 2. An average of 400,000,000 reads per sample was obtained (approximately 80,000 reads per cell).

Alignment, barcode assignment and unique molecular identifier counting. The Cell Ranger 1.3 Single-Cell software suite (10X Genomics) was implemented to process the raw sequencing data from the Illumina HiSeq run. This pipeline performed demultiplexing, alignment (using the mouse genome mm10 from ENSEMBL GRCm38) and barcode processing to generate gene–cell matrices used for downstream analysis. Specifically, data from two control-sgRNA- and two *Regnase-1*-sgRNA-transduced TIL OT-I cell samples were combined into one dataset for consistent filtering, and unique molecular identifiers (UMIs) mapped to genes encoding ribosomal proteins were removed. Cells with low UMI counts (potentially dead cells with broken membranes) or high UMI counts (potentially two or more cells in a single droplet) were filtered. A small fraction of outlier cells (888) was further removed because of their low transcriptome diversity (meaning that fewer genes were detected than in other cells with a comparable number of captured UMIs). A total of 13,879 cells (control-sgRNA-transduced, 6,811; *Regnase-1*-sgRNA-transduced, 7,068) were captured, with an average of 11,040 mRNA molecules (UMIs, median: 9,391; range: 2,928–44,330). We normalized the expression level of each gene to 100,000 UMIs per cell and log-transformed them by adding 0.5 to the expression matrix.

Data visualization. Underlying cell variations derived from control-sgRNA- and *Regnase-1*-sgRNA-transduced TIL OT-I cell single-cell gene-expression data were visualized in a two-dimensional projection by *t*-SNE. Expression of individual genes or pathway scores was colour-coded (from low to high, blue–red) for each cell on *t*-SNE plots. To visualize *Tcf7*-expressing cells, we defined *Tcf7*^{high} cells as cells with the highest third quantile of *Tcf7* expression (with \log_2 (gene expression intensity) = 2.910317 as threshold) among all cells.

Statistical analysis for biological experiments

For biological experiment (non-omics) analyses, data were analysed using Prism 6 software (GraphPad) by two-tailed paired Student's *t*-test, two-tailed unpaired Student's *t*-test, or one-way ANOVA with Newman–Keuls's test. Two-way ANOVA was performed for comparing tumour

growth curves. The log-rank (Mantel–Cox) test was performed for comparing mouse survival curves. $P < 0.05$ was considered significant. Data are presented as mean \pm s.d. or mean \pm s.e.m.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Microarray, RNA-seq, ATAC-seq and scRNA-seq data have been deposited in the NCBI Gene Expression Omnibus (GEO) database and are accessible through the GEO SuperSeries accession number GSE126072. Source Data for Figs. 1–5 and Extended Data Figs. 1–9 are provided with the paper. All other relevant data are available from the corresponding author upon reasonable request.

40. Platt, R. J. et al. CRISPR–Cas9 knockin mice for genome editing and cancer modeling. *Cell* **159**, 440–455 (2014).
41. Hogquist, K. A. et al. T cell receptor antagonist peptides induce positive selection. *Cell* **76**, 17–27 (1994).
42. Overwijk, W. W. et al. Tumor regression and autoimmunity after reversal of a functionally tolerant state of self-reactive CD8⁺ T cells. *J. Exp. Med.* **198**, 569–580 (2003).
43. Chen, R. et al. In vivo RNA interference screens identify regulators of antiviral CD4⁺ and CD8⁺ T cell differentiation. *Immunity* **41**, 325–338 (2014).
44. Birsoy, K. et al. An essential role of the mitochondrial electron transport chain in cell proliferation is to enable aspartate synthesis. *Cell* **162**, 540–551 (2015).
45. Sanson, K. R. et al. Optimized libraries for CRISPR–Cas9 genetic screens with multiple modalities. *Nat. Commun.* **9**, 5416 (2018).
46. Sukumar, M. et al. Mitochondrial membrane potential identifies cells with enhanced stemness for cellular therapy. *Cell Metab.* **23**, 63–76 (2016).
47. Wang, W. et al. Effector T cells abrogate stroma-mediated chemoresistance in ovarian cancer. *Cell* **165**, 1092–1105 (2016).
48. Wei, J. et al. Autophagy enforces functional integrity of regulatory T cells by coupling environmental cues and metabolic homeostasis. *Nat. Immunol.* **17**, 277–285 (2016).
49. Zeng, H. et al. mTORC1 couples immune signals and metabolic programming to establish T_{reg} cell function. *Nature* **499**, 485–490 (2013).
50. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
51. Karmaus, P. W. F. et al. Metabolic heterogeneity underlies reciprocal fates of T_H17 cell stemness and plasticity. *Nature* **565**, 101–105 (2019).
52. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
53. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
54. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
55. Zhang, Y. et al. Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, R137 (2008).
56. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
57. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
58. Calvo, S. E., Clauser, K. R. & Mootha, V. K. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res.* **44**, D1251–D1257 (2016).
59. Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
60. Ramirez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
61. Cuellar-Partida, G. et al. Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics* **28**, 56–62 (2012).
62. Krishnamoorthy, V. et al. The IRF4 gene regulatory module functions as a read-write integrator to dynamically coordinate T helper cell fate. *Immunity* **47**, 481–497 (2017).

Acknowledgements The authors acknowledge M. Hendren for animal colony management, C. Li for help with plasmids, G. Neale and S. Olsen for assistance with sequencing and St. Jude Immunology FACS core facility for cell sorting. This work was supported by NIH AI105887, AI131703, AI140761, AI150241, AI150514, and CA221290 (to H.C.).

Author contributions J.W. conceived the project, designed and performed in vitro and in vivo experiments, analysed data and wrote the manuscript; L.L. performed molecular experiments and analysed data; W.Z. performed CAR T-cell-related experiments and analysed data, with guidance from T.L.G., who also provided CAR transgenic mice and human CD19-Ph⁺ B-ALL cell line; Y.D. performed bioinformatic analyses; S.A.L. helped to perform cellular experiments; C.G. performed imaging experiments; Y.W. performed Seahorse experiments; Y.-D.W. and J.Y. analysed CRISPR–Cas9 screening data; C.Q. performed scRNA-seq data analyses, with guidance from J.Y.; B.X. helped with ATAC-seq analysis; A.K. helped with molecular cloning; J.S. helped with ATAC-seq sample preparation; H.H. helped to perform scRNA-seq experiments; J.G.D. designed and generated the lentiviral sgRNA metabolic library and provided guidance for CRISPR–Cas9 screening data analyses; and H.C. helped to conceive and design experiments, co-wrote the manuscript and provided overall direction.

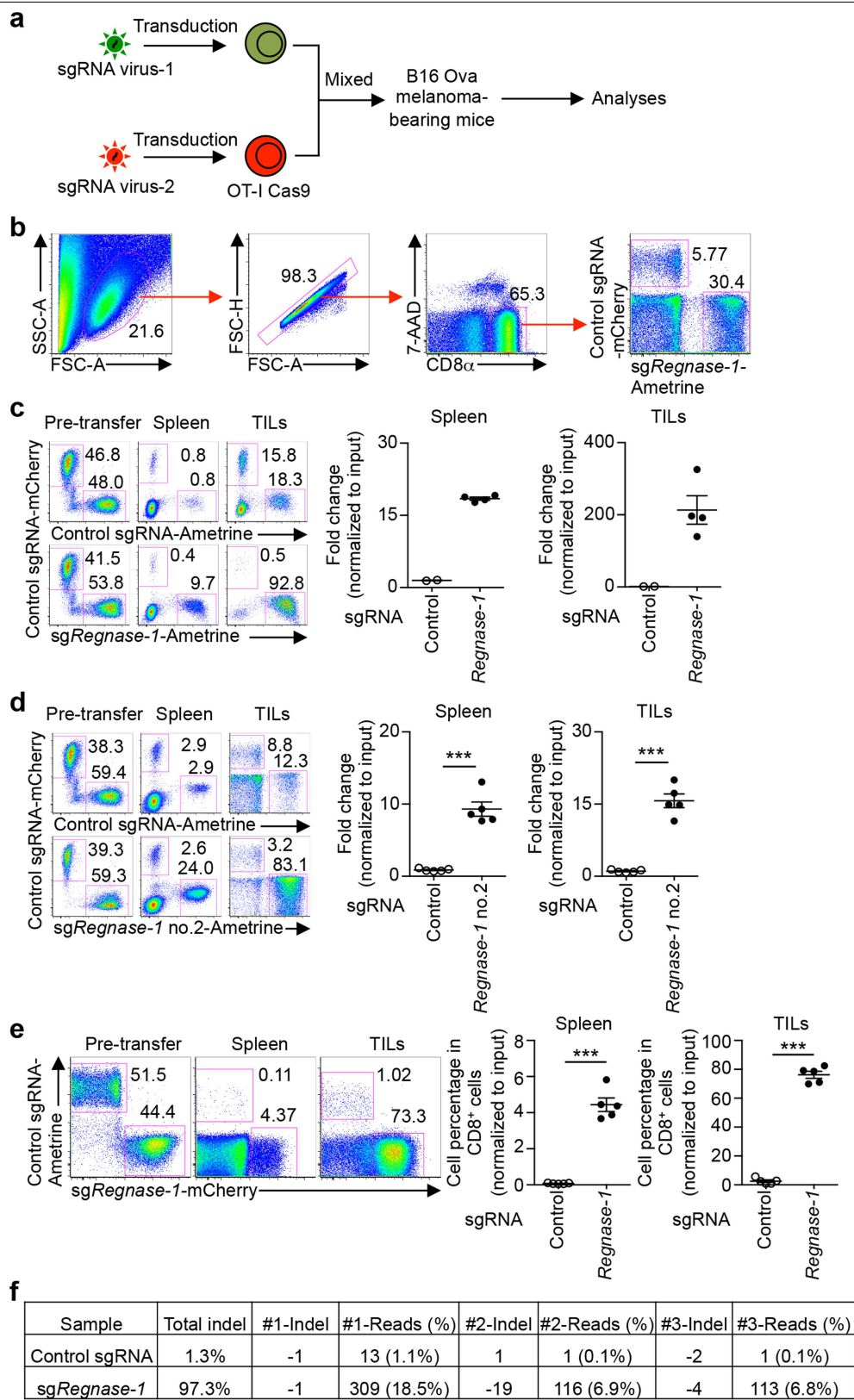
Competing interests H.C. and J.W. are authors of a patent application related to REGNASE-1 and BATF.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1821-z>.

Correspondence and requests for materials should be addressed to H.C.

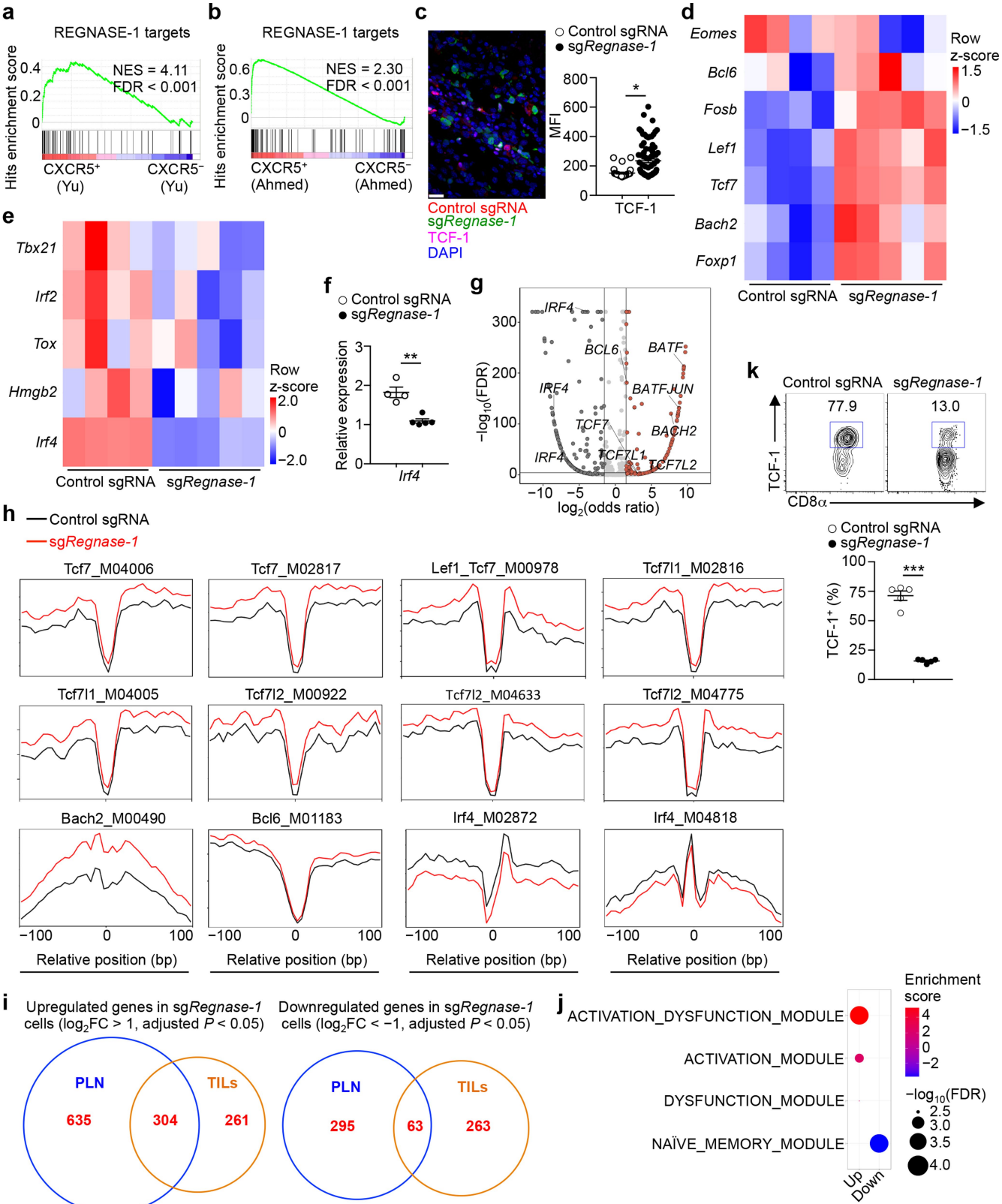
Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1|See next page for caption.

Extended Data Fig. 1 | Validation of the effect of REGNASE-1 deletion on CD8⁺ T cell accumulation in tumour immunity using the in vivo dual transfer system. **a**, Diagram of the in vivo dual transfer system. OT-I cells transduced with sgRNA viral vectors expressing distinct fluorescent proteins were mixed and transferred into the same tumour-bearing hosts, in which further analyses were performed. **b**, Gating strategy for sgRNA-transduced OT-I cell analysis. **c**, **d**, OT-I cells transduced with non-targeting control sgRNA (mCherry⁺) were mixed at a 1:1 ratio with cells either transduced with control sgRNA (ametrine⁺) (**c** (*n* = 2), **d** (*n* = 5), left, top) or two different sgRNAs targeting *Regnase-1* (*Regnase-1* sgRNA, ametrine⁺, **c** (*n* = 4), left, bottom; or *Regnase-1* sgRNA no. 2, ametrine⁺, **d** (*n* = 5), left, bottom), and transferred into tumour-bearing hosts. Mice were analysed at 7 days after adoptive transfer for the proportion of OT-I cells in CD8 α ⁺ cells (**c**, **d**, left), and the quantification of relative OT-I cell percentages in CD8 α ⁺ cells (normalized to input) in the spleen and TILs

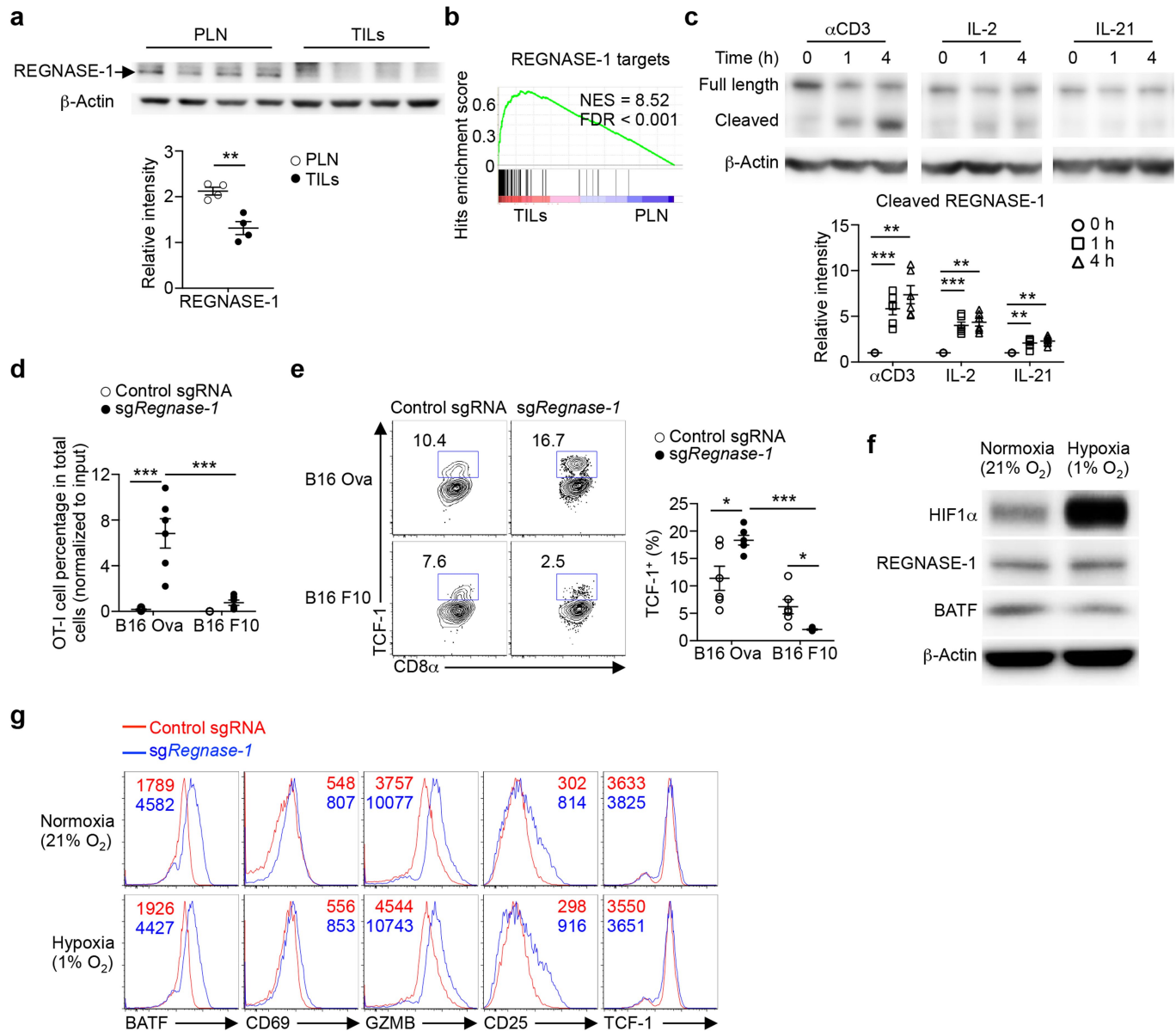
(**c**, **d**, right). Numbers in plots indicate the frequencies of OT-I cells. **e**, OT-I cells transduced with control sgRNA (ametrine⁺) were mixed at a 1:1 ratio with cells transduced with *Regnase-1* sgRNA (mCherry⁺) and transferred into tumour-bearing hosts (*n* = 5). Mice were analysed at 7 days after adoptive transfer for the proportion of OT-I cells in CD8 α ⁺ cells (left), and the quantification of relative OT-I cell percentage in CD8 α ⁺ cells (normalized to input) in the spleen and TILs (right). Numbers in plots indicate the frequencies of OT-I cells. **f**, Insertion and deletion (indel) mutations after CRISPR targeted disruption in OT-I cells transduced with either control sgRNA or *Regnase-1* sgRNA, via deep sequencing analysis of indels generated at the exonic target site of the *Regnase-1* gene, including 97.3% of indel events in *Regnase-1*-sgRNA-transduced cells isolated from tumours compared to 1.3% in control-sgRNA-transduced cells. Mean \pm s.e.m. (**c–e**). ****P* < 0.001. Two-tailed unpaired Student's *t*-test (**d**, **e**). Data are representative of two independent experiments (**e**).



Extended Data Fig. 2 | See next page for caption.

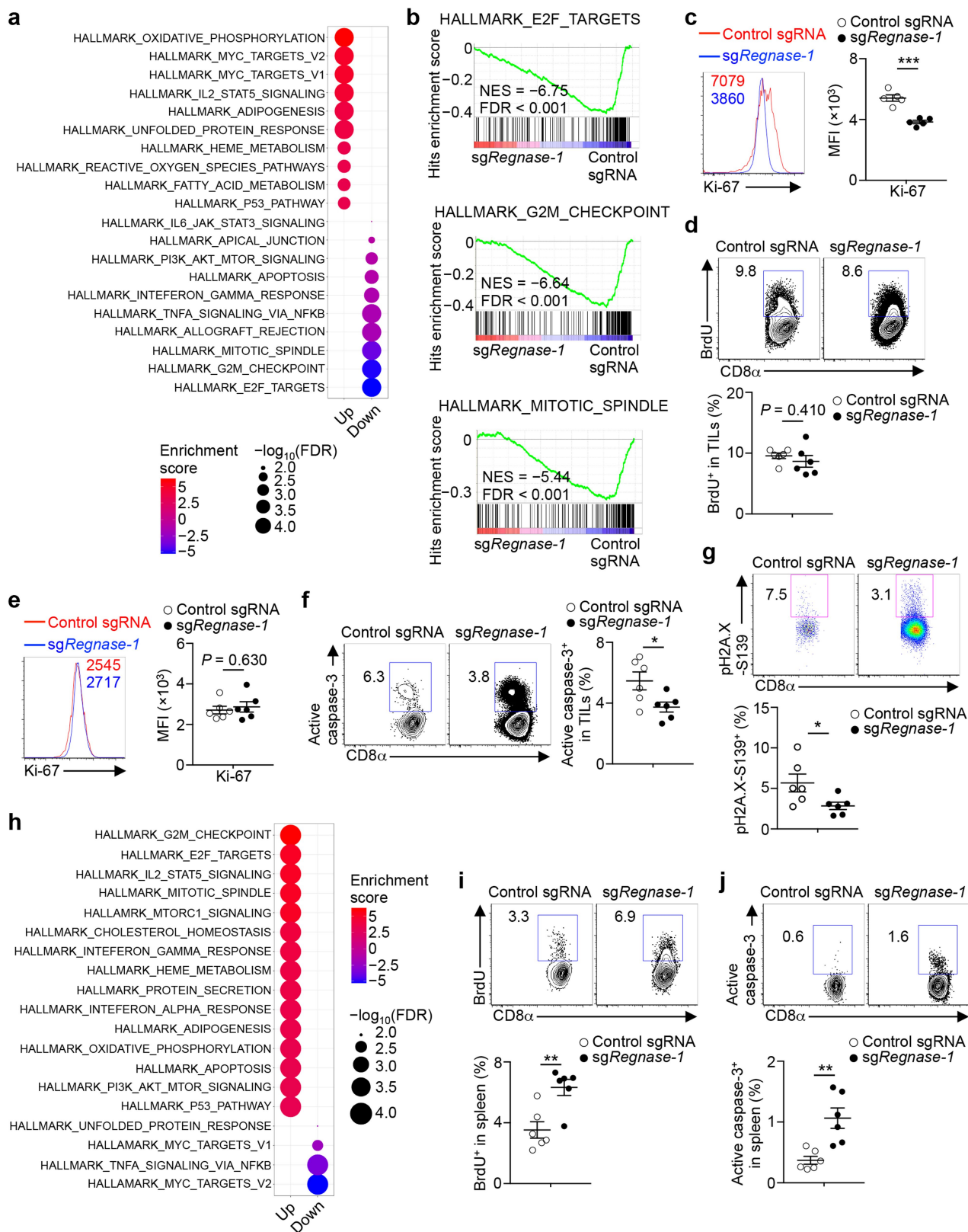
Extended Data Fig. 2 | Tumour-infiltrating and peripheral REGNASE-1-null CD8⁺ T cells show distinct immune signatures. **a, b**, GSEA enrichment plots of antigen-specific CXCR5⁺ and CXCR5⁻ exhausted CD8⁺ T cells from chronic infection using gene targets repressed by REGNASE-1 (that is, the top 100 upregulated genes in TIL *Regnase-1*-sgRNA- compared to control-sgRNA-transduced OT-I cells, as identified using RNA-seq). **c**, Representative images (left) and quantification of MFI (right) of TCF-1 expression (pink) in control-sgRNA- (mCherry⁺; red) and *Regnase-1*-sgRNA-transduced OT-I cells (ametrine⁺; green) in the whole-tumour section ($n = 4$ mice). Scale bars, 20 μ m. **d, e**, Gene-expression heat maps normalized by row (z-score) for the naive- or memory-T-cell-associated transcription factors (**d**) or effector- or exhausted-T-cell-associated transcription factors (**e**) in control-sgRNA- ($n = 4$) and *Regnase-1*-sgRNA ($n = 5$)-transduced OT-I cells isolated from TILs. Specifically, control-sgRNA- and *Regnase-1*-sgRNA-transduced OT-I cells were mixed and transferred into tumour-bearing mice, and tumour-infiltrating OT-I cells were isolated at day 7 for transcriptional profiling by RNA-seq. **f**, Real-time PCR analysis of *Irf4* mRNA expression in control-sgRNA- ($n = 4$ samples) and *Regnase-1*-sgRNA ($n = 5$ samples)-transduced OT-I cells isolated from TILs. **g**, Summary of ATAC-seq motif enrichment data showing \log_2 (odds ratio) and $-\log_{10}$ (FDR) of cells from control-sgRNA- and *Regnase-1*-sgRNA-transduced OT-I cells isolated from TILs ($n = 4$ samples per group). Specifically, control-sgRNA- and *Regnase-1*-sgRNA-transduced OT-I cells were mixed and transferred into tumour-bearing mice, and tumour-infiltrating OT-I cells were isolated at day 7 for ATAC-seq analysis. **h**, Tn5 insert sites from ATAC-seq analysis were aligned to motifs for

transcription factors from the TRANSFAC database, and the binding profiles of TCF-1, BACH2, BCL6 and IRF4 are shown. **i**, Venn diagram showing the overlap of significantly upregulated (left, *Regnase-1*-sgRNA- ($n = 5$ samples) versus control-sgRNA-transduced OT-I cells ($n = 4$ samples)) or downregulated genes (right, *Regnase-1*-sgRNA- versus control-sgRNA-transduced OT-I cells) by RNA-seq profiling between TIL and PLN OT-I cells. Specifically, control-sgRNA- and *Regnase-1*-sgRNA-transduced OT-I cells were mixed and transferred into tumour-bearing mice, and OT-I cells were isolated at day 7 for transcriptional profiling by RNA-seq. **j**, GSEA enrichment plots of PLN *Regnase-1*-sgRNA- ($n = 5$) versus control-sgRNA ($n = 4$)-transduced OT-I cells using gene sets of four different tumour-infiltrating CD8 T cell activation states¹¹. Specifically, control-sgRNA- and *Regnase-1*-sgRNA-transduced OT-I cells were mixed and transferred into tumour-bearing mice, and PLN OT-I cells were isolated at day 7 for transcriptional profiling by RNA-seq. **k**, OT-I cells transduced with control sgRNA (mCherry⁺) and *Regnase-1* sgRNA (ametrine⁺) were mixed and transferred into tumour-bearing mice ($n = 5$ mice), and OT-I cells in the spleen were analysed at day 7 for expression of TCF-1 (top), and quantification of the frequency of TCF-1⁺ cells (bottom). Numbers in graphs indicate the frequencies of cells in gates. Mean \pm s.e.m. (**c, f, k**). * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Kolmogorov–Smirnov test followed by Benjamini–Hochberg correction (**a, b, j**), two-tailed unpaired Student's *t*-test (**c, f, k**), two-sided Fisher's exact test followed by Benjamini–Hochberg correction (**g**) or two-sided Fisher's exact test (**i**). Data are representative of two independent experiments (**c, f, k**).



Extended Data Fig. 3 | Upstream signals regulate REGNASE-1 expression and REGNASE-1-null cell phenotypes. **a**, Immunoblot analysis of REGNASE-1 expression in control-sgRNA-transduced OT-I cells isolated from PLN and TILs at 7 days after adoptive transfer ($n = 4$ samples per group) (top). Quantification of the relative intensity of REGNASE-1 expression (bottom). β -Actin is used as a loading control. **b**, GSEA enrichment plots of PLN and TIL control-sgRNA-transduced OT-I cells ($n = 4$) used in **a**, by using gene targets repressed by REGNASE-1 (that is, the top 100 upregulated genes in TIL *Regnase-1*-sgRNA-compared to control-sgRNA-transduced cells, as identified using RNA-seq). **c**, OT-I cells were stimulated with anti-CD3 and anti-CD28 overnight before viral transduction, and then cultured in IL-7- and IL-15-containing medium for another 3 days in vitro. Pre-activated OT-I cells were then stimulated with anti-CD3, IL-2 or IL-21 for 0, 1 and 4 h ($n = 5$ samples per group) for immunoblot analysis of full-length and cleaved REGNASE-1 (top), and quantification of the relative intensity of cleaved REGNASE-1 expression (bottom). β -Actin is used as a loading control. **d**, **e**, OT-I cells transduced with control sgRNA (mCherry⁺) and *Regnase-1* sgRNA (ametrine⁺) were mixed at a 1:1 ratio and transferred into

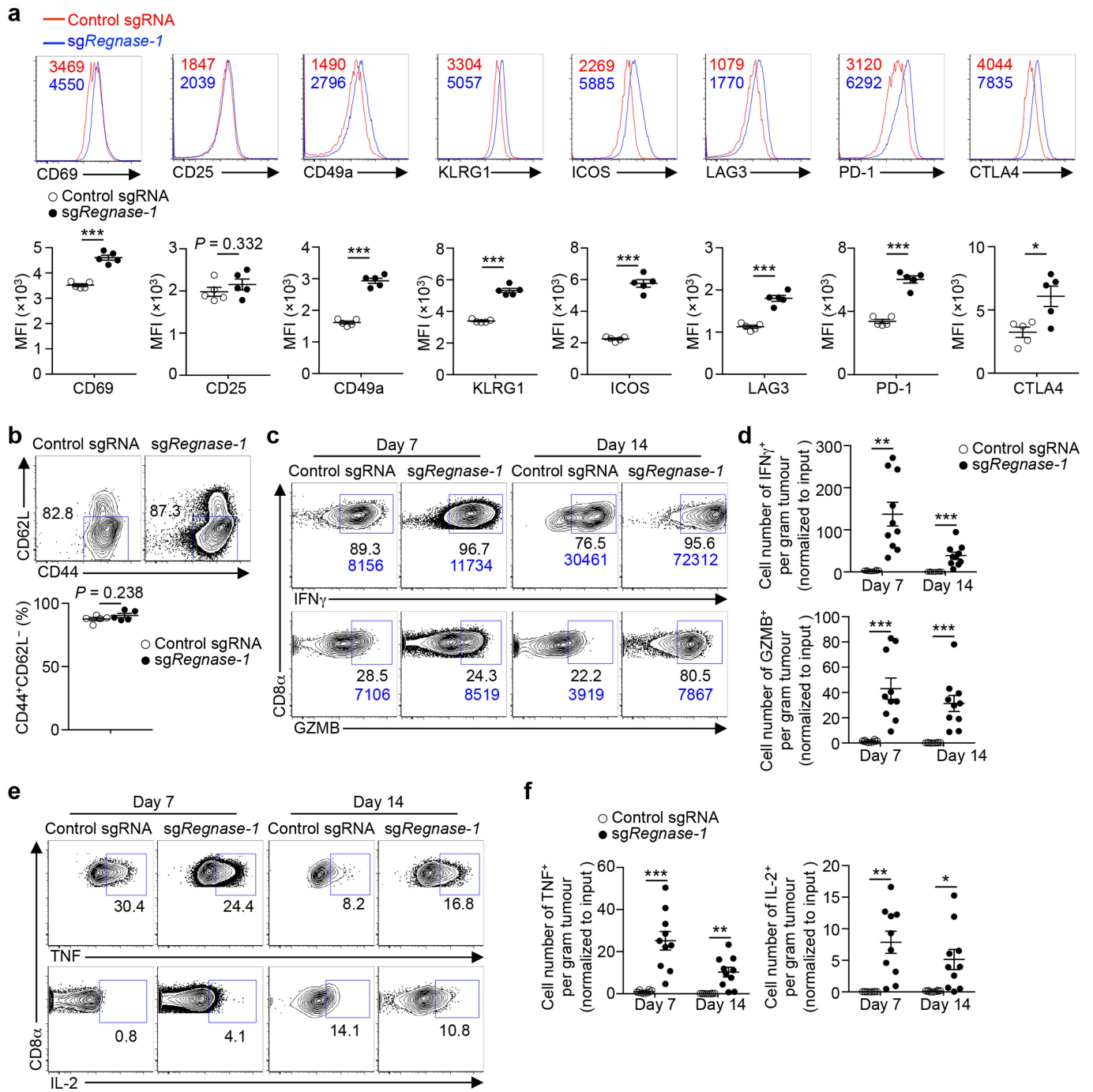
mice bearing B16 Ova ($n = 6$ mice) or B16 F10 ($n = 6$ mice) tumours. Mice were analysed at day 7 after adoptive transfer for quantification of relative OT-I cell percentage in total cells (normalized to input) in the TILs (**d**) and expression of TCF-1 (**e**, left), and quantification of the frequency of TCF-1⁺ cells (**e**, right) in tumour-infiltrating OT-I cells. **f**, **g**, OT-I cells were stimulated with anti-CD3 and anti-CD28 overnight before viral transduction, and then cultured in IL-2-, IL-7- and IL-15-containing medium for another 3 days in vitro. Pre-activated OT-I cells were then continuously cultured in normoxia (21% O₂) or hypoxia (1% O₂) conditions for 48 h for immunoblot analysis of expression of HIF1 α , REGNASE-1 and BATF (**f**), and for flow cytometry analysis of the expression of BATF, CD69, GZMB, CD25 and TCF-1 (**g**). Numbers in graphs indicate MFI (**g**). β -Actin is used as a loading control. Mean \pm s.e.m. (**a**, **c**–**e**). * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Two-tailed unpaired Student's *t*-test (**a**), Kolmogorov–Smirnov test followed by Benjamini–Hochberg correction (**b**) or one-way ANOVA (**c**–**e**). Data are representative of two (**c**, **f**, **g**) independent experiments, or pooled from two (**d**, **e**) independent experiments.



Extended Data Fig. 4 | See next page for caption.

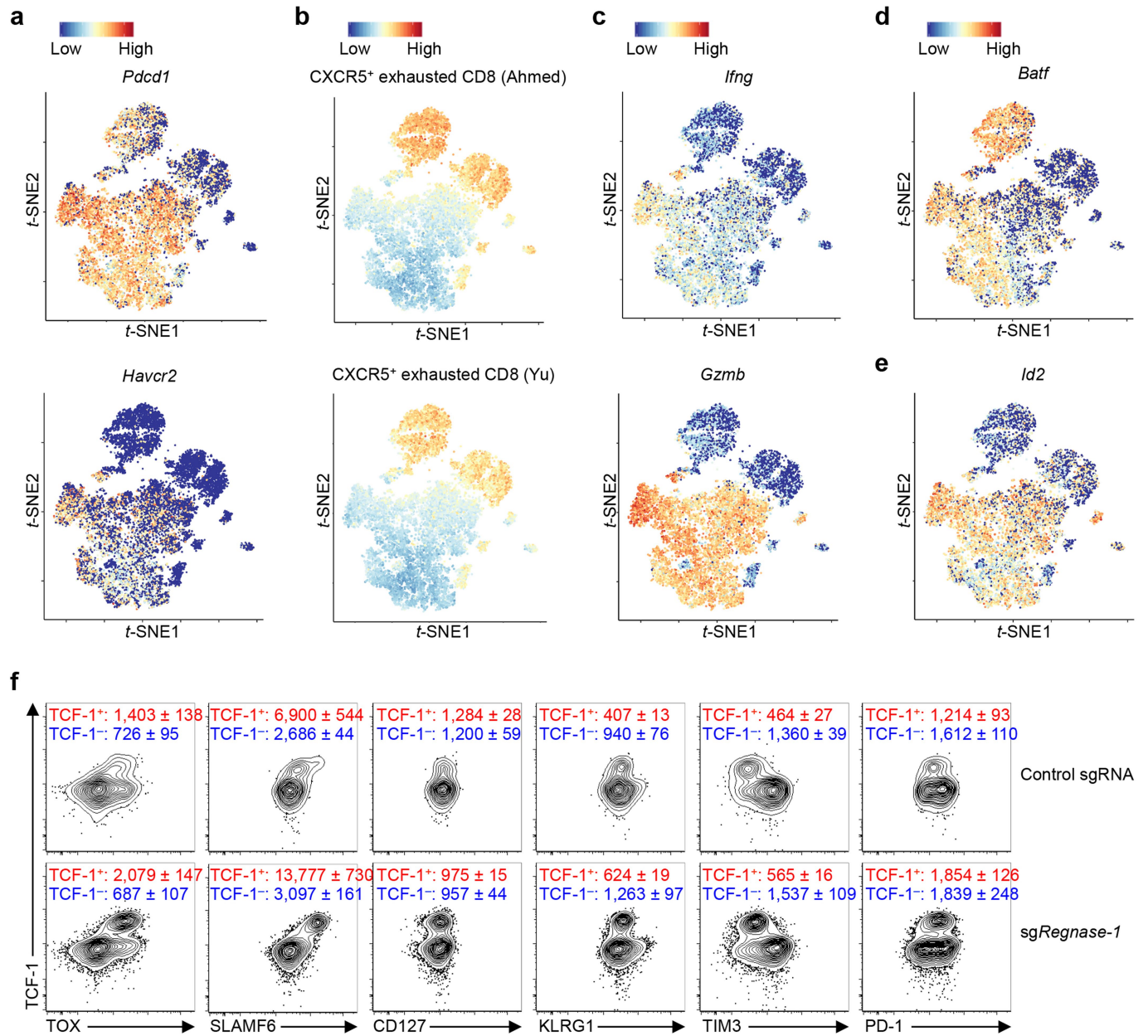
Extended Data Fig. 4 | Proliferation and survival analyses of REGNASE-1-null CD8⁺ T cells in tumour immunity. **a**, List of the top-10 significantly (FDR < 0.05) upregulated and downregulated pathways in TIL *Regnase-1*-sgRNA-transduced OT-I cells, as revealed by performing GSEA using Hallmark gene sets. Specifically, control-sgRNA- (*n* = 4) and *Regnase-1*-sgRNA (*n* = 5)-transduced OT-I cells were mixed and transferred into tumour-bearing mice, and tumour-infiltrating OT-I cells were isolated at day 7 for transcriptional profiling by RNA-seq. **b**, GSEA enrichment plots of TIL *Regnase-1*-sgRNA-transduced OT-I cells using cell-cycling-associated gene sets, including E2F targets (top), G2M checkpoint (middle) and mitotic spindle (bottom). **c–g**, OT-I cells transduced with control sgRNA (mCherry⁺) and *Regnase-1* sgRNA (ametrine⁺) were mixed and transferred into tumour-bearing mice, and tumour-infiltrating OT-I cells were analysed at day 7 (**d–g**) (*n* = 6 mice) and day 14 (**c**) (*n* = 5 mice) by flow cytometry for Ki-67 expression (**c**, left; **e**, left), BrdU incorporation (**d**, top; pulse for 18 h), active caspase-3 expression (**f**, left), Ser139 phosphorylation of histone variant H2A.X (**g**, top), and quantification of MFI of Ki-67 (**c**, right; **e**, right), frequency of BrdU⁺ cells (**d**, bottom), frequency of active caspase-3⁺ cells (**f**, right) and the frequency of the Ser139-phosphorylated histone variant H2A.X⁺ cells (**g**, bottom). Numbers in graphs indicate the MFI of Ki-67 (**c**, left; **e**, left). Numbers in plots indicate the frequencies of BrdU⁺ cells (**d**, top),

active caspase-3⁺ cells (**f**, left) and Ser139-phosphorylated histone variant H2A.X⁺ cells (**g**, top). **h**, List of the top-15 significantly (FDR < 0.05) upregulated and top-4 significantly downregulated pathways in PLN *Regnase-1*-sgRNA-transduced OT-I cells, as revealed by performing GSEA using Hallmark gene sets. Specifically, control-sgRNA- (*n* = 4) and *Regnase-1*-sgRNA (*n* = 5)-transduced OT-I cells were mixed and transferred into tumour-bearing mice, and PLN OT-I cells were isolated at day 7 for transcriptional profiling by RNA-seq. **i, j**, OT-I cells transduced with control sgRNA (mCherry⁺) and *Regnase-1* sgRNA (ametrine⁺) were mixed and transferred into tumour-bearing mice, and OT-I cells in the spleen were analysed at day 7 (**i, j**) (*n* = 6 mice) by flow cytometry for BrdU incorporation (**i**, top; pulse for 18 h) and active caspase-3 expression (**j**, top), and quantification of frequencies of BrdU⁺ cells (**i**, bottom) and active caspase-3⁺ cells (**j**, bottom). Numbers in plots indicate the frequencies of BrdU⁺ cells (**i**, top) and active caspase-3⁺ cells (**j**, top). Mean ± s.e.m. (**c–g, i, j**). **P* < 0.05, ***P* < 0.01, ****P* < 0.001. Kolmogorov–Smirnov test followed by Benjamini–Hochberg correction (**a, b, h**) or two-tailed unpaired Student's *t*-test (**c–g, i, j**). Data are representative of two (**c**) independent experiments, or pooled from two (**d–g, i, j**) independent experiments.



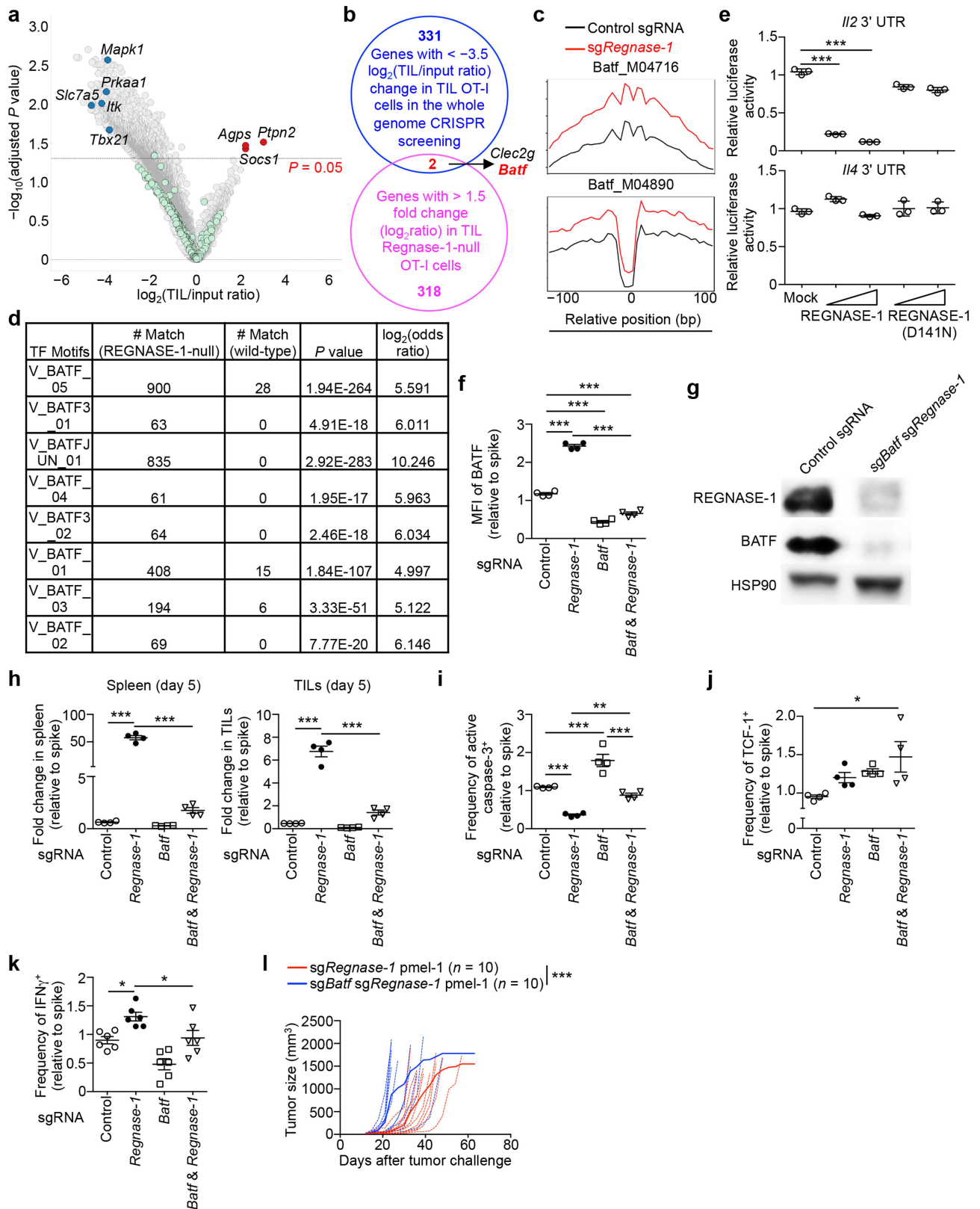
Extended Data Fig. 5 | Effector molecule expression of tumour-infiltrating REGNASE-1-null CD8⁺ T cells. **a, b**, OT-I cells transduced with control sgRNA (mCherry⁺) or *Regnase-1* sgRNA (ametrine⁺) were mixed at a 5:1 ratio and transferred into tumour-bearing mice ($n = 5$ mice), and tumour-infiltrating OT-I cells were analysed at day 7 for the expression of CD69, CD25, CD49a, KLRG1, ICOS, LAG3, PD-1 and CTLA4 (**a**, top) and CD44 and CD62L (**b**, top), and quantification of MFI of CD69, CD25, CD49a, KLRG1, ICOS, LAG3, PD-1 and CTLA4 (**a**, bottom) and frequency of CD44⁺CD62L⁻ cells (**b**, bottom). The numbers in graphs indicate the MFI (**a**, top). The numbers in plots indicate the frequency of CD44⁺CD62L⁻ cells (**b**, top). **c–f**, OT-I cells transduced with control sgRNA (mCherry⁺) or *Regnase-1* sgRNA (ametrine⁺) were mixed at a 5:1 ratio and transferred into tumour-bearing mice, and analysed at day 7 ($n = 10$ mice) or

day 14 ($n = 10$ mice). Flow cytometry analysis of expression of IFN γ (**c**, top), GZMB (**c**, bottom), TNF (**e**, top) and IL-2 (**e**, bottom) in TIL OT-I cells, and quantification of the numbers of IFN γ ⁺ cells (**d**, top), GZMB⁺ cells (**d**, bottom), TNF⁺ cells (**f**, left) and IL-2⁺ cells (**f**, right) per gram of tumour (normalized to input). The numbers adjacent to outlined areas indicate the frequencies of IFN γ ⁺ cells and the MFI of IFN γ in IFN γ ⁺ cells (**c**, top), and the frequency of GZMB⁺ cells and the MFI of GZMB in GZMB⁺ cells (**c**, bottom), and the frequencies of TNF⁺ cells (**e**, top) or IL-2⁺ cells (**e**, bottom). Mean \pm s.e.m. (**a, b, d, f**). * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Two-tailed unpaired Student's *t*-test (**a, b**) or two-tailed paired Student's *t*-test (**d, f**). Data are representative of two (**a–c, e**) independent experiments, or pooled from two (**d, f**) independent experiments.



Extended Data Fig. 6 | scRNA-seq and flow cytometry analyses of tumour-infiltrating REGNASE-1-null OT-I cells. **a–e**, scRNA-seq analysis of control-sgRNA- and *Regnase-1*-sgRNA-transduced OT-I cells isolated from TILs. Specifically, control-sgRNA- and *Regnase-1*-sgRNA-transduced OT-I cells were mixed and transferred into tumour-bearing mice, and tumour-infiltrating OT-I cells were isolated at day 7 for transcriptional profiling by scRNA-seq. t-SNE visualization of *Pdcd1* (**a**, top), *Havcr2* (**a**, bottom), *Ifng* (**c**, top), *Gzmb* (**c**, bottom), *Batf* (**d**) and *Id2* (**e**) gene expression, and ‘CXCR5⁺ exhausted CD8 (Ahmed)¹²’ (**b**, top) and ‘CXCR5⁺ exhausted CD8 (Yu)¹³’ (**b**, bottom) gene

signatures in individual cells. **f**, OT-I cells transduced with control sgRNA and *Regnase-1* sgRNA were mixed and transferred into tumour-bearing mice ($n = 5$ mice; data from 1 representative mouse are shown), and tumour-infiltrating OT-I cells were analysed at day 7 for the expression of TOX, SLAMF6, CD127, KLRG1, TIM3 and PD-1 in TCF-1⁺ and TCF-1⁻ cells of control-sgRNA- and *Regnase-1*-sgRNA-transduced OT-I cells. Numbers in graphs indicate the mean \pm s.e.m. of MFI of markers on the x-axis after gating on TCF-1⁺ or TCF-1⁻ subsets. Data are representative of two independent experiments (**f**).

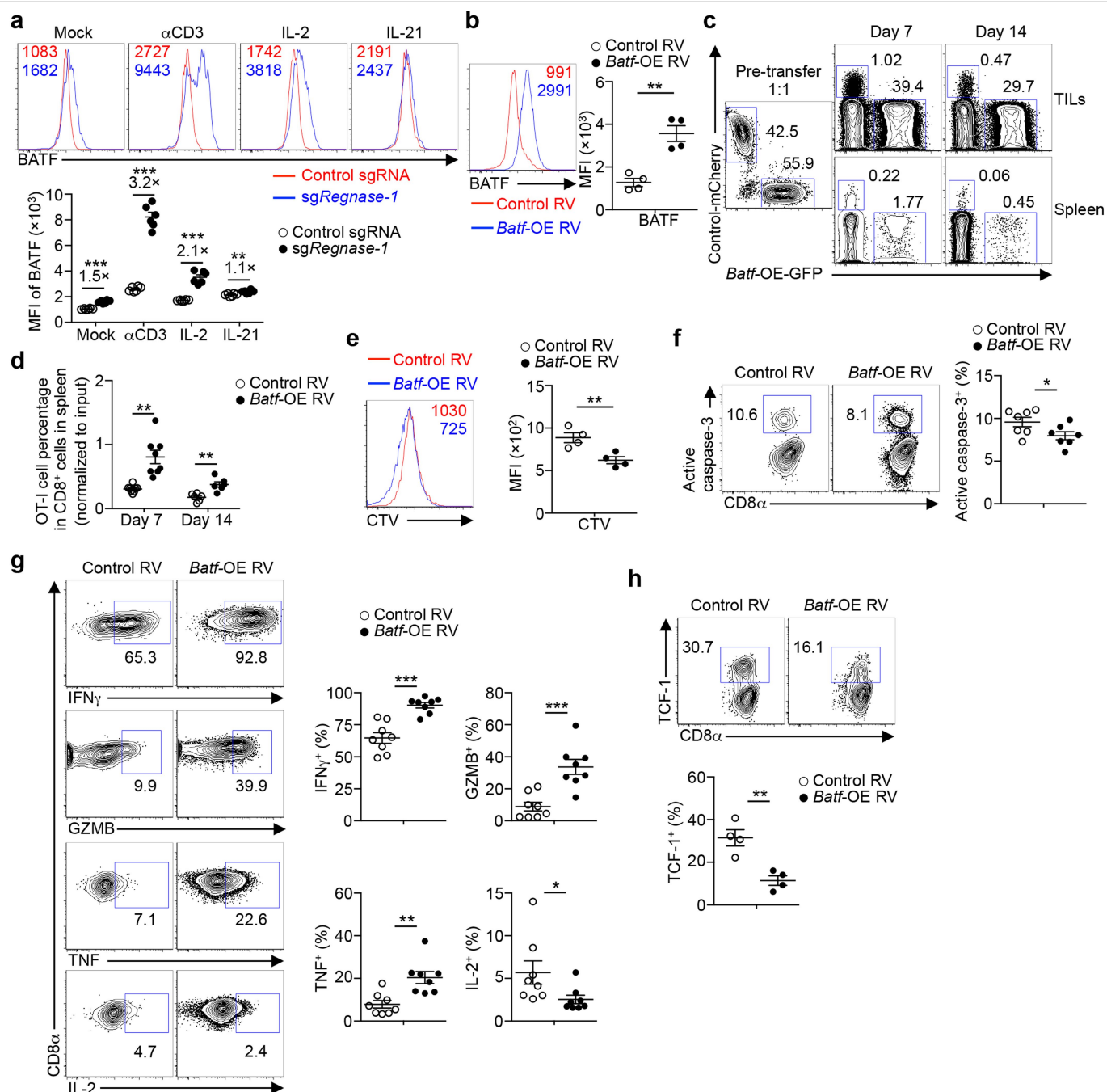


Extended Data Fig. 7 | See next page for caption.

Article

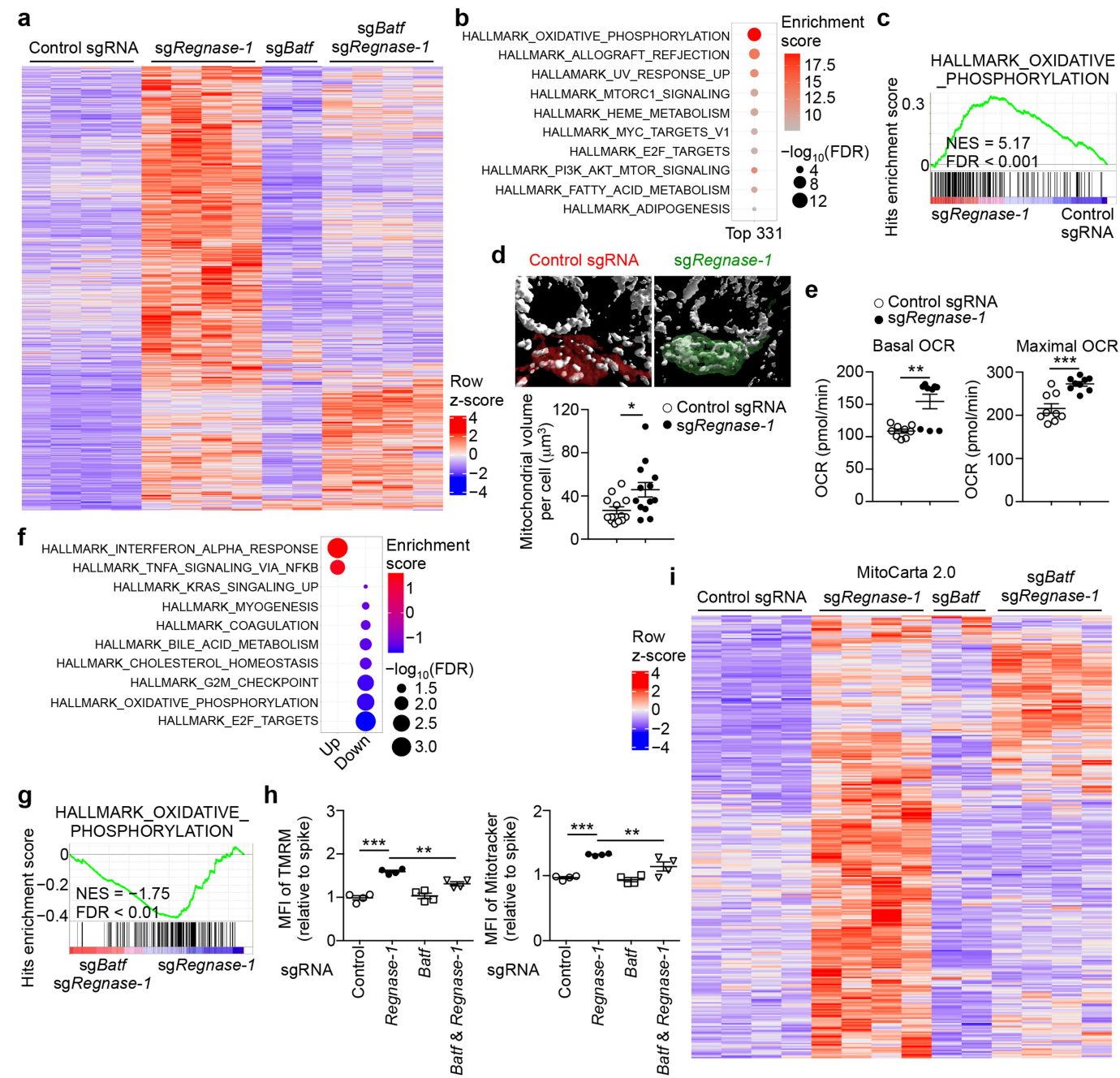
Extended Data Fig. 7 | Genome-scale CRISPR screening identifies BATF as an important REGNASE-1 functional target in tumour immunity. **a**, Scatter plot of the enrichment of each gene versus its adjusted P value in genome-scale CRISPR screening. Gene enrichment was calculated by averaging the enrichment of the corresponding sgRNAs ($n = 4$ for each gene) in tumour-infiltrating OT-I cells relative to input ($\log_2(\text{TIL}/\text{input ratio})$), with the most extensively enriched (red) and selectively depleted (blue) genes (adjusted $P < 0.05$), as well as dummy genes (green, generated by random combinations of 4 out of 1,000 non-targeting control sgRNAs per dummy gene). **b**, Venn diagram showing the overlap of genes between top-depleted genes in genome-scale CRISPR screening (by less than $-3.5 \log_2(\text{TIL}/\text{input ratio})$; adjusted $P < 0.05$) and top-upregulated genes in TIL *Regnase-1*-sgRNA- versus control-sgRNA-transduced OT-I cells as identified by RNA-seq (by greater than 1.5 fold change (\log_2 -transformed ratio); adjusted $P < 0.05$). **c**, Tn5 insert sites from ATAC-seq analysis were aligned to motifs for transcription factors from the TRANSFAC database, and the binding profiles of BATF are shown. **d**, Enrichment of BATF-binding motifs in the genomic regions with upregulated accessibility in REGNASE-1-null cells. First, we analysed common regions in our REGNASE-1-null ATAC-seq data and published BATF ChIP-seq peaks (GSE54191²⁶). Next, we scanned these common regions with TRANSFAC motifs for BATF, and numbers of motif matches and associated Fisher's exact test P values and $\log_2(\text{odds ratios})$ are shown (a positive $\log_2(\text{odds ratio})$ value indicates that a motif is more likely to occur in REGNASE-1-null cells than in wild-type samples; 'E-x' denotes ' $\times 10^{-x}$ '). **e**, Luciferase activity of HEK293T cells measured at 48 h after transfection with *Il2* mRNA 3' UTR (top) or *Il4* mRNA 3' UTR (bottom) luciferase reporter plasmid, together with control (mock), wild-type REGNASE-1- or REGNASE-1(D141N)-expressing plasmid ($n = 3$ samples per

group). **f**, OT-I cells transduced with control sgRNA (mCherry⁺; spike) were mixed at a 1:1 ratio with cells transduced with control sgRNA (ametrine⁺), *Regnase-1* sgRNA (ametrine⁺), *Batf* sgRNA (GFP⁺) or *Batf* and *Regnase-1* sgRNAs (GFP⁺ and ametrine⁺), and transferred into tumour-bearing hosts individually ($n = 4$ mice per group). Mice were analysed at 5 days after adoptive transfer for quantification of relative MFI of BATF normalized to spike in the tumour-infiltrating OT-I cells (**f**). **g**, Immunoblot analysis of REGNASE-1 and BATF expression in in vitro cultured OT-I cells 3 days after transduction with control sgRNA or *Batf* and *Regnase-1* sgRNAs. HSP90 is used as a loading control. **h-k**, The same transfer system as in **f** was used. Five days after adoptive transfer, mice were analysed for the quantification of relative OT-I cell percentage in CD8 α^+ cells normalized to spike in the spleen (**h**, left, $n = 4$) and TILs (**h**, right, $n = 4$). Tumour-infiltrating OT-I cells were analysed at day 5 ($n = 4$ mice per group) for the quantification of the relative frequency of active caspase-3⁺ cells normalized to spike (**i**), and the quantification of the relative frequency of TCF-1⁺ cells normalized to spike (**j**), or at day 7 ($n = 6$ mice per group) for quantification of the relative frequency of IFN γ^+ cells normalized to spike (**k**). **l**, Four million pmel-1 cells transduced with *Regnase-1* sgRNA (ametrine⁺) ($n = 10$ recipients) or *Batf* and *Regnase-1* sgRNAs (GFP⁺ and ametrine⁺) ($n = 10$ recipients) were transferred into mice at day 12 after B16 F10 melanoma engraftment, followed by analysis of tumour size. Mean \pm s.d. (**e**). Mean \pm s.e.m. (**f**, **h-k**). * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Two-tailed unpaired Student's paired t -test followed by Bonferroni correction (**a**), two-sided Fisher's exact test (**d**), one-way ANOVA (**e**, **f**, **h-k**) or two-way ANOVA (**l**). Data are representative of two (**e**) or three (**g**) independent experiments, or pooled from two (**f**, **h-l**) independent experiments.



Extended Data Fig. 8 | BATF overexpression markedly enhances CD8 $^+$ T cell antitumour responses. **a**, OT-I cells were stimulated with anti-CD3 and anti-CD28 overnight before viral transduction, and then cultured in IL-7 and IL-15-containing medium for another 3 days in vitro. Control-sgRNA- and *Regnase-1*-sgRNA-transduced OT-I cells were then stimulated with anti-CD3, IL-2 or IL-21 overnight for flow cytometry analysis of BATF expression (top), and quantification of the MFI of BATF (bottom) ($n = 6$ samples per group). Numbers in graphs indicate the MFI (top) and fold change between comparisons (bottom). **b–h**, OT-I cells transduced with control retrovirus (RV; mCherry $^+$) were mixed at a 1:1 ratio with cells transduced with *Batf*-overexpressing retrovirus (GFP $^+$), and transferred into tumour-bearing hosts. Mice were analysed at day 4 (**e**) ($n = 4$ mice), day 5 (**b, h**) ($n = 4$ mice), day 7 (**c, d, f, g**) ($n = 6–8$ mice) or day 14 (**c, d**) ($n = 6$ mice) for the expression of BATF (**b**, left), active caspase-3 (**f**, left), IFN γ , GZMB, TNF and IL-2 (**g**, left) and TCF-1 (**h**, top) in

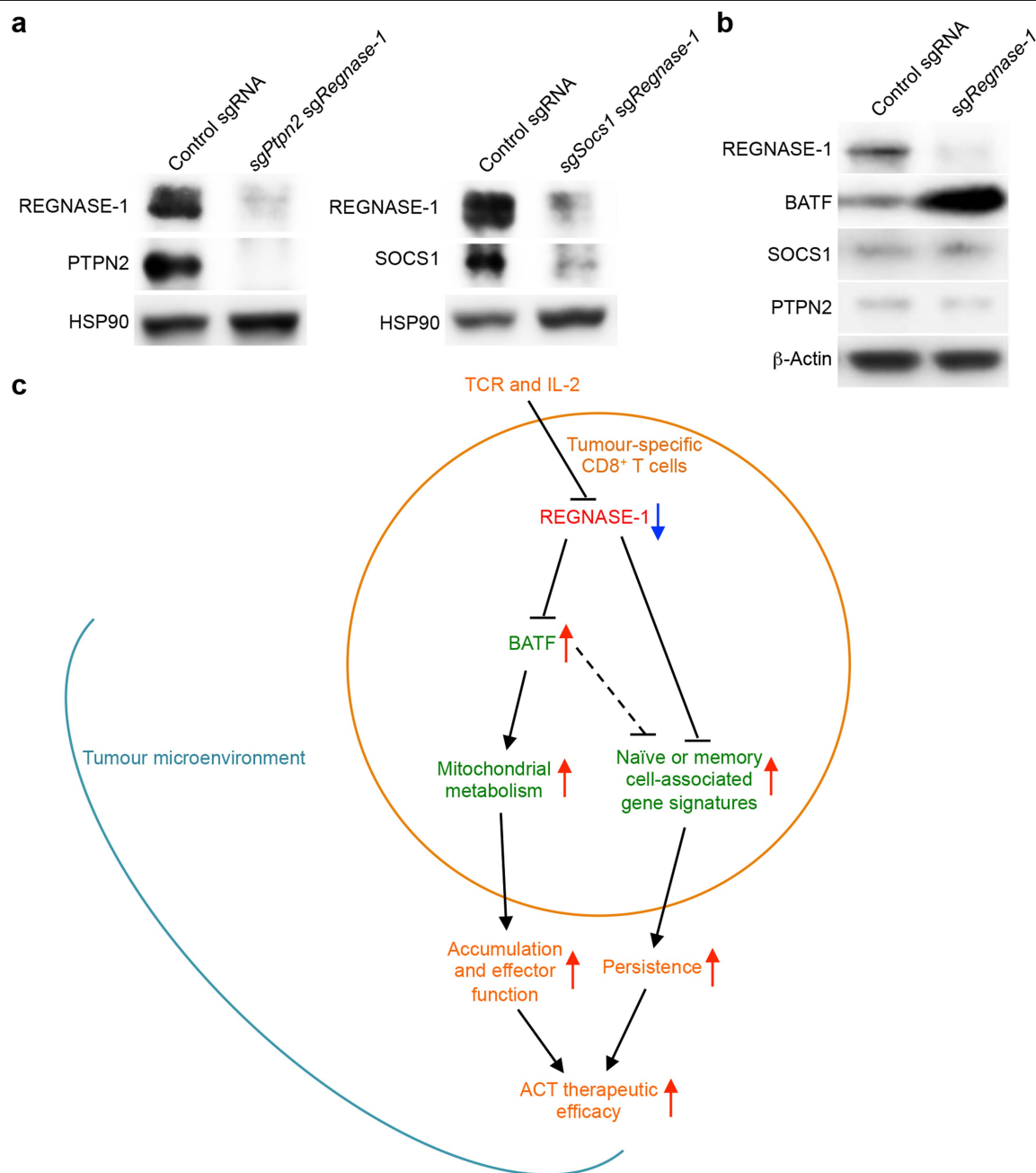
TIL OT-I cells; the quantification of the MFI of BATF in TIL OT-I cells (**b**, right); quantification of the frequencies of active caspase-3 $^+$ cells (**f**, right), IFN γ $^+$, GZMB $^+$, TNF $^+$ and IL-2 $^+$ cells (**g**, right) and TCF-1 $^+$ cells (**h**, bottom) in TIL OT-I cells; analysis of the proportion of donor-derived OT-I cells in total CD8 $^+$ cells in TILs and spleen (**c**); the quantification of the relative OT-I cell percentage in CD8 $^+$ cells in the spleen (normalized to input) (**d**); the dilution of CellTrace Violet (CTV) in TIL OT-I cells (**e**, left); and the quantification of the MFI of CTV in TIL OT-I cells (**e**, right). The numbers in graphs indicate the MFI (**b**, left; **e**, left), the frequencies of OT-I cells in gates (**c**), the frequency of active caspase-3 $^+$ cells (**f**, left), the frequencies of IFN γ $^+$, GZMB $^+$, TNF $^+$ or IL-2 $^+$ cells (**g**, left), and the frequency of TCF-1 $^+$ cells (**h**, top). Mean \pm s.e.m. (**a, b, d–h**). * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Two-tailed unpaired Student's t -test (**a, b, d–h**). Data are representative of two (**a, c**) independent experiments, or pooled from two (**b, d–h**) independent experiments.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Genome-scale CRISPR screening identifies mitochondrial metabolism as an important downstream pathway of REGNASE-1 and BATF. **a**, Chromatin accessibility heat maps normalized by row (z-score) for 7,480 genes with significantly increased chromatin accessibility (by |fold change (\log_2 -transformed ratio)| > 0.5; $P < 0.05$) in *Regnase-1*-sgRNA-transduced OT-I cells as compared to control-sgRNA-transduced cells. Specifically, OT-I cells transduced with control sgRNA (mCherry⁺) ($n = 4$), *Regnase-1* sgRNA (ametrine⁺) ($n = 4$), *Batf* sgRNA (GFP⁺) ($n = 2$) or *Batf* and *Regnase-1* sgRNAs (GFP⁺ and ametrine⁺) ($n = 4$) were transferred into tumour-bearing hosts individually. OT-I cells were isolated from TILs at day 7 for ATAC-seq analysis. We annotated the differential accessibility regions in ATAC-seq for the nearest genes, and identified 7,480 genes with significantly increased chromatin accessibility in REGNASE-1-null cells as compared to wild-type cells. BATF co-deletion reversed the upregulated chromatin accessibility for a large proportion of these genes (5,052 in total). Also, 2,527 of these 5,052 genes showed significantly downregulated chromatin accessibility in BATF-null cells as compared to wild-type cells. **b**, Functional enrichment plots of the top-10 significantly (FDR < 0.05) enriched pathways in top-ranking depleted genes ($n = 4$ sgRNAs for each gene) identified in the genome-scale CRISPR screening (by less than $-3.5 \log_2$ (TIL/input ratio); adjusted $P < 0.05$). **c**, GSEA enrichment plots of TIL *Regnase-1*-sgRNA-transduced OT-I cells using the OXPHOS Hallmark gene set. Specifically, control-sgRNA- and *Regnase-1*-sgRNA-transduced OT-I cells were mixed and transferred into tumour-bearing mice, and tumour-infiltrating OT-I cells were isolated at day 7 for transcriptional profiling by RNA-seq. **d**, Representative images (top) and quantification of mitochondrial volume (stained with TOM20, white) per cell (bottom) in control-sgRNA- (mCherry⁺; red) and *Regnase-1*-sgRNA-transduced OT-I cells (ametrine⁺; green) in tumours at 7 days after adoptive transfer ($n = 4$ mice). **e**, Oxygen consumption rate (OCR) bioenergetic profiling of control-sgRNA- and *Regnase-1*-sgRNA-transduced OT-I cells cultured in vitro for basal (left) and maximal (right) OCR ($n = 9$ samples per group). **f**, List of the top-2 significantly (FDR < 0.05) upregulated and top-8 significantly downregulated pathways in

TIL *Batf*- and *Regnase-1*-sgRNAs- versus *Regnase-1*-sgRNA-transduced OT-I cells ($n = 3$ samples per group) isolated from TILs, as revealed by performing GSEA using the Hallmark gene sets. Specifically, *Regnase-1*-sgRNA- and *Batf*- and *Regnase-1*-sgRNA-transduced OT-I cells were mixed and transferred into tumour-bearing mice, and tumour-infiltrating OT-I cells were isolated at day 7 for transcriptional profiling by microarray. **g**, GSEA enrichment plots of TIL *Batf*- and *Regnase-1*-sgRNAs- versus *Regnase-1*-sgRNA-transduced OT-I cells ($n = 3$ samples per group) using the OXPHOS gene set. **h**, OT-I cells transduced with control sgRNA (mCherry⁺; spike) were mixed at a 1:1 ratio with cells transduced with control sgRNA (ametrine⁺), *Regnase-1* sgRNA (ametrine⁺), *Batf* sgRNA (GFP⁺) or *Batf* and *Regnase-1* sgRNA (GFP⁺ and ametrine⁺), and transferred into tumour-bearing hosts individually ($n = 4$ mice per group). Mice were analysed at 5 days after adoptive transfer for quantification of the relative MFI of TMRM (left) and Mitotracker (right), normalized to spike in tumour-infiltrating OT-I cells. **i**, Chromatin accessibility heat maps normalized by row (z-score) for mitochondrial genes with significantly increased chromatin accessibility (by |fold change (\log_2 -transformed ratio)| > 0.5; $P < 0.05$) in *Regnase-1*-sgRNA-transduced OT-I cells compared to control-sgRNA-transduced cells, determined by ATAC-seq as described in **a**. We annotated the differential accessibility regions in ATAC-seq for the nearest genes, and superimposed these genes with 1,158 mitochondrial genes defined in the MitoCarta 2.0 database. A total of 341 mitochondrial genes showed significantly upregulated chromatin accessibility in the absence of REGNASE-1, 214 of which were blocked by BATF co-deletion in BATF-null REGNASE-1-null cells. Moreover, 96 of these 214 genes showed significantly downregulated chromatin accessibility in BATF-null cells as compared to wild-type cells. Mean \pm s.e.m. (**d**, **e**, **h**). * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Two-sided Fisher's exact test (**a**, **i**), right-tailed Fisher's exact test (**b**), Kolmogorov-Smirnov test followed by Benjamini-Hochberg correction (**c**, **f**, **g**), two-tailed unpaired Student's *t*-test (**d**, **e**) or one-way ANOVA (**h**). Data are representative of two (**d**, **e**) independent experiments, or pooled from two (**h**) independent experiments.



Extended Data Fig. 10 | Targeting PTPN2 and SOCS1 and model of REGNASE-1 functions in tumour-specific CD8⁺ T cells. **a**, Immunoblot analysis of REGNASE-1, PTPN2 and SOCS1 expression in OT-1 cells cultured in vitro, 3 days after transduction with control sgRNA, *Ptpn2* and *Regnase-1* sgRNAs (left) or *Socs1* and *Regnase-1* sgRNAs (right). HSP90 is used as a loading control. **b**, Immunoblot analysis of REGNASE-1, BATF, SOCS1 and PTPN2 expression in OT-1 cells transduced with control sgRNA or *Regnase-1* sgRNA, cultured in vitro for 3 days after viral transduction. β-Actin is used as a loading control. **c**, REGNASE-1 is a major negative regulator of CD8⁺ T cell antitumour responses, and TCR and IL-2 inhibit its expression and activity. Deletion of REGNASE-1 unleashes a potent therapeutic efficacy of engineered tumour-specific CD8⁺

T cells against cancers, by coordinating transcriptional and metabolic programs to achieve greatly improved cell accumulation and function. As a key functional target of REGNASE-1, excessive BATF drives robust cell accumulation and effector function—in part through enhancing mitochondrial metabolism—in REGNASE-1-null CD8⁺ T cells. REGNASE-1 deletion also reprograms cells to acquire increased gene signatures associated with naïve or memory cells and to gain survival advantage, which contribute to the improved persistence of REGNASE-1-null effector CD8⁺ T cells. Targeting PTPN2 and SOCS1 (not depicted here) acts in coordination with REGNASE-1 inhibition to promote CD8⁺ T cell antitumour responses. Data are representative of three independent experiments (**a**, **b**).

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

No software was used.

Data analysis

Flowjo 9.9.4 (Tree Star) for FACS results; GraphPad Prism 6 for statistics; CLC Genomics Workbench v11 (Qiagen) for Hi-seq FastQ file analysis; Trimmomatic v.0.36, star v.2.5.2b and R package DEseq2 v. 1.18.1 for RNA-seq analysis; Affymetrix Expression Console v.1.1, R package limma v.3.34.9, R package ggplot2 v.2.2.1 and limma v.3.34.9 for micro-array analysis; BWA version 0.7.16, Picard version 2.9.4, samtools version 1.9, IGV version 2.4.13, MACS2 (version 2.1.1.20160309, bedtools v2.24.0, voom package (R 3.23, edgeR 3.12.1, limma 3.26.9), deeptools v2.5.7, MEME suite version 4.9.0 and MEME suite version 4.11.3 for ATAC-seq analysis; NIS Elements software (Nikon Instruments) and Slidebook software (Intelligent Imaging Innovations) for imaging data analysis; The Cell Ranger 1.3 Single-Cell software suite (10x Genomics) and t-distributed stochastic neighbour embedding (tSNE) for scRNA-seq analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Microarray, RNA-seq, ATAC-seq and scRNA-seq data have been deposited in the NCBI Gene Expression Omnibus (GEO) database and are accessible through the GEO SuperSeries accession number: GSE126072. All other relevant data are available from the corresponding author upon reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was selected to maximize the chance of uncovering mean difference which is also statistically significant.
Data exclusions	No data were excluded.
Replication	All the experimental findings were reliably reproduced as validated by at least two independent experiments.
Randomization	In tumour therapy experiments, at day 12 after tumour inoculation, mice bearing similar size of tumours were randomly divided into groups.
Blinding	Blinding was not relevant to these studies.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used

The following antibodies were used for flow cytometry: Active caspase-3 staining was performed using instructions and reagents from the Active Caspase-3 Apoptosis Kit (BD Biosciences). BrdU staining (pulsed for 18 h) was performed using instructions and reagents from the APC BrdU Flow Kit (BD Biosciences). 7-AAD (Sigma) or fixable viability dye (eBioscience) was used for dead-cell exclusion. anti-IFN γ (XMG1.2), anti-TNF (MAB11), anti-IL-2 (JES6-5H4), anti-CD69 (H1.2F3), anti-CD25 (PC61.5), anti-KLRG1 (2F1), anti-ICOS (7E.17G9), anti-LAG3 (C9B7W), anti-PD-1 (J43), anti-CTLA4 (1B8), anti-TOX (TXRX10), anti-TIM3 (RMT3-23) (all from eBioscience); anti-GZMB (QA16A02), anti-CD49a (HMA1), anti-CD44 (IM7), anti-Ki-67 (16A8), anti-CD127 (A7R34) (all from Biolegend); anti-BrdU (3D4), anti-active caspase-3 (C92-605), anti-pH2A.X-S139 (N1-431) (DNA damage biomarker, which measures phosphorylation of the histone variant H2A.X at Ser13946,47), anti-SLAMF6 (13G3) (all from BD Biosciences); anti-BATF (D7C5), anti-TCF-1 (C63D9) (all from Cell Signaling Technology); anti-CD8 α (53-6.7) (from SONY); anti-CD62L (MEL-14) (from TONBO Bioscience). 1:100–1:200 dilution.

The following antibodies were used for western blot: anti-MCPIP1 antibody (604421) (R&D), anti-BATF (D7C5) (Cell Signaling Technology), anti-PTPN2 (E-11) (Santa Cruz Biotechnology), anti-SOCS1 (E-9) (Santa Cruz Biotechnology), anti-Hsp90 (MAB3286) (R&D) and anti- β -actin (8H10D10) (Cell Signaling Technology), and HRP conjugated anti-mouse IgG (W4021) (from Promega). 1:1000 – 1:5000 dilutions.

The following antibodies were used for imaging: anti-mCherry (Biorbyt orb11618), anti-GFP (Rockland Immuno 600-401-215), anti-TCF-7 (C63D9) (Cell Signaling Technology 2203), and anti-Tom20 (2F8.1) (Millipore MABT166). 1:100–1:200 dilution.

Validation

The specificities of listed FACS antibodies have been validated by the manufacturer by flow cytometry.
The specificities of listed WB antibodies have been validated by the manufacturer by western blot.
The specificities of listed imaging antibodies have been validated by the manufacturer by imaging.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	B16 F10 and HEK293T cell lines were purchased from ATCC. B16 Ova cell line was kindly provided by Dr. Dario Vignali. huCD19-Ph+ B-ALL cell line was provided by Dr. Terrence Geiger.
Authentication	Cell lines were not authenticated.
Mycoplasma contamination	Cell lines were not tested for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	C57BL/6, OT-I, pmel-1 and Rosa26-Cas9 knockin mice were purchased from The Jackson Laboratory. CAR-T transgenic mice were provided by Terrence Geiger (to be described elsewhere). We crossed Rosa26-Cas9 knockin mice with OT-I, pmel-1 or CAR-T transgenic mice to express Cas9 in antigen-specific CD8+ T cells. Gender-matched mice were used at 7–16 weeks old unless otherwise noted.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve samples collected from the field.
Ethics oversight	Experiments and procedures were performed in accordance with the Institutional Animal Care and Use Committee (IACUC) of St. Jude Children's Research Hospital.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	The spleens and peripheral lymph nodes (PLNs) were gently separated under nylon mesh using the flat end of a 3-mL syringes. Red blood cells were removed by ACK lysing buffer, followed by washing cells with isolation buffer. After spinning down, the cell pellets were resuspended and filtered with nylon mesh before staining. For the examination of tumour infiltrating lymphocytes, tumours were excised, minced and digested with 0.5 mg/ml Collagenase IV (Roche) + 200 U/ml DNase I (Sigma) for 1 h at 37 °C, and then passed through 70-µm filters to remove undigested tumour tissues. TILs were then isolated by density-gradient centrifugation over Percoll.
Instrument	LSRII or LSR Fortessa (BD Biosciences)
Software	Flowjo 9.9.4 or later (Tree Star)
Cell population abundance	The purities of the sorted sgRNA transduced cells were more than 98%.
Gating strategy	Based on the pattern of FSC-A/SSC-A, cells in the lymphocyte gate were used for analysis of T cell subsets. Singlets were gated according to the pattern of FSC-H vs. FSC-A. Positive populations were determined by the specific antibodies, which were distinct from negative populations.

- ☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Plasma membrane V-ATPase controls oncogenic RAS-induced macropinocytosis

<https://doi.org/10.1038/s41586-019-1831-x>

Craig Ramirez^{1,2}, Andrew D. Hauser^{1,2}, Emily A. Vucic¹ & Dafna Bar-Sagi^{1*}

Received: 4 January 2018

Accepted: 21 October 2019

Published online: 11 December 2019

Oncogenic activation of RAS is associated with the acquisition of a unique set of metabolic dependencies that contribute to tumour cell fitness. Cells that express oncogenic RAS are able to internalize and degrade extracellular protein via a fluid-phase uptake mechanism termed macropinocytosis¹. There is increasing recognition of the role of this RAS-dependent process in the generation of free amino acids that can be used to support tumour cell growth under nutrient-limiting conditions². However, little is known about the molecular steps that mediate the induction of macropinocytosis by oncogenic RAS. Here we identify vacuolar ATPase (V-ATPase) as an essential regulator of RAS-induced macropinocytosis. Oncogenic RAS promotes the translocation of V-ATPase from intracellular membranes to the plasma membrane via a pathway that requires the activation of protein kinase A by a bicarbonate-dependent soluble adenylate cyclase. Accumulation of V-ATPase at the plasma membrane is necessary for the cholesterol-dependent plasma-membrane association of RAC1, a prerequisite for the stimulation of membrane ruffling and macropinocytosis. These observations establish a link between V-ATPase trafficking and nutrient supply by macropinocytosis that could be exploited to curtail the metabolic adaptation capacity of RAS-mutant tumour cells.

To identify essential mediators of RAS-driven macropinocytosis, we conducted a full genome short interfering RNA (siRNA) screen using a microscopy-based high-throughput assay in which oncogenic HRAS (HRAS(G12V))-dependent induction of macropinocytosis in HeLa cells is measured by uptake of fluorescently labelled high-molecular-mass dextran³. Confirmed hits from the screen displaying more than 70% inhibition of macropinocytosis were analysed using STRING (<http://string-db.org/>). Four main networks emerged from this analysis, corresponding to splicing, actin, ubiquitination and V-ATPase (Fig. 1a, Extended Data Fig. 1a). Given the high enrichment of hits mapping to the V-ATPase protein complex and the increasing appreciation for the role of V-ATPase in tumorigenesis and metastasis⁴, we focused on delineating the functional link between V-ATPase and oncogenic RAS-induced macropinocytosis.

V-ATPase is required for macropinocytosis

V-ATPase is a multi-subunit transmembrane complex that transduces protons from the cytoplasm to the lumen of organelles and across the plasma membrane⁴. The role of V-ATPase in regulating a wide array of membrane trafficking and intracellular transport processes is well documented; however, its contribution to macropinocytosis has not been described⁴. We confirmed that knockdown of V-ATPase subunits identified in the screen inhibit macropinocytosis in HeLa HRAS(G12V) cells (Extended Data Fig. 1b). Among these, knockdown of the ATP6V1A (V1A) subunit, a component of the catalytic domain of the pump, was accompanied by the strongest inhibitory effect on macropinocytosis,

and was thus selected for further analyses. Knockdown of V1A using siRNA targeting the 3' untranslated region (UTR) led to a 90% inhibition of macropinocytosis, which could be rescued by ectopic expression of a 3×Flag-tagged V1A construct, ruling out off-target effects (Fig. 1b, c, Extended Data Fig. 1c).

V-ATPase is required for trafficking of cholesterol from endocytic organelles to the plasma membrane⁵. Given that RAC1, a critical regulator of macropinocytosis, is dependent on plasma membrane cholesterol for its localization and activation^{6,7}, we tested whether the contribution of V-ATPase to macropinocytosis might be causally related to cholesterol-dependent plasma membrane localization of RAC1. Consistent with previous results^{5,8}, siRNA-mediated depletion of V-ATPase resulted in loss of cholesterol from the plasma membrane and its accumulation in intracellular punctae (Fig. 1d–f, Extended Data Fig. 1d). No change in overall cholesterol level was detected following depletion of V-ATPase (Extended Data Fig. 1e). Additionally, treatment with the V-ATPase inhibitor bafilomycin A1 resulted in loss of plasma membrane cholesterol, consistent with a requirement for V-ATPase activity in cholesterol trafficking (Extended Data Fig. 1f). Fluorescence microscopy and biochemical analyses of RAC1 localization demonstrated that redistribution of cholesterol in cells in which V-ATPase was depleted or inhibited was accompanied by reduced levels of plasma membrane-associated RAC1 (Fig. 1d, g, Extended Data Fig. 1f, g). Plasma membrane localization of RAC1 has been linked to its activation state⁶. However, we observed no change in RAC1–GTP levels following V-ATPase depletion (Extended Data Fig. 2a, b) indicating that dependence of RAC1 on V-ATPase for membrane localization is uncoupled from changes

¹Department of Biochemistry and Molecular Pharmacology, New York University School of Medicine, New York, NY, USA. ²These authors contributed equally: Craig Ramirez, Andrew D. Hauser.

*e-mail: dafna.bar-sagi@nyulangone.org

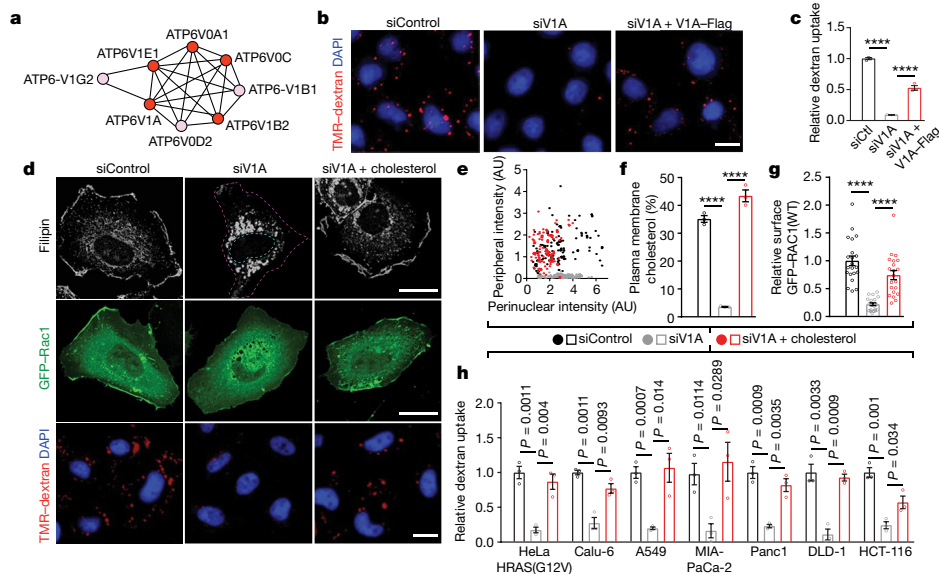


Fig. 1 | V-ATPase is required for RAS-induced macropinocytosis and plasma membrane-localized cholesterol. **a**, V-ATPase cluster defined by STRING analysis (pink, primary screen; red, primary and confirmation screen). **b**, **c**, Effect of V-ATPase depletion using siRNA directed against the V1A subunit (siV1A) and rescue (siV1A + V1A-Flag) on macropinocytosis in HeLa T7-HRAS(G12V) cells. **b**, Fluorescence micrographs showing tetramethylrhodamine (TMR)-dextran uptake. **c**, Quantification of TMR-dextran uptake. **d–g**, Effect of V-ATPase depletion on cholesterol distribution, RAC1 localization and macropinocytosis in HeLa HRAS(G12V) cells treated as shown. **d**, Fluorescence micrographs showing cholesterol localization (filipin,

top), GFP-RAC1 localization (middle) and TMR-dextran uptake (bottom). Dashed lines delineate the cell and nucleus. **e**, **f**, Quantification of cholesterol distribution displayed as scatter plot (**e**; each dot represents a cell) and bar graph (**f**). **g**, Quantification of plasma membrane localization of GFP-RAC1. **h**, Quantification of cholesterol-dependent dextran uptake in mutant RAS cells. Images (**b**, **d**) are representative of three biological replicates. Scale bars, 10 μ m. At least 500 (**c**, **h**), 50 (**e**, **f**) and 7 (**g**) cells were quantified in each biological replicate ($n = 3$); data are mean \pm s.e.m.; unpaired two-tailed Student's *t*-test; **** $P < 0.0001$.

in RAC1 activity, a conclusion further validated using a constitutively active mutant of RAC1, GFP-RAC1(L61). Similarly to wild-type RAC1, this mutant failed to interact with the plasma membrane in cells depleted of V-ATPase despite being fully active (Extended Data Fig. 2a, c).

Addition of exogenous cholesterol is a commonly used approach to deliver cholesterol to membranes⁸. We found that addition of cholesterol to growth medium of V-ATPase-depleted or V-ATPase-inhibited cells restored plasma membrane pools of cholesterol and rescued macropinocytosis and RAC1 membrane localization (Fig. 1d–h, Extended Data Figs. 1f, 2d). By contrast, addition of cholesterol to RAC1-depleted cells failed to rescue macropinocytosis (Extended Data Fig. 2e). Furthermore, introduction of a RAC1 construct in which the native cholesterol-dependent membrane-targeting sequence was replaced with the cholesterol-independent membrane-targeting sequence of KRAS into V-ATPase-depleted cells was sufficient to rescue macropinocytosis⁹ (Extended Data Fig. 2c). Lastly, the requirement of V-ATPase for macropinocytosis, which could be bypassed by the addition of cholesterol or the cholesterol-independent RAC1 construct, was displayed by multiple mutant KRAS cell lines (Fig. 1h, Extended Data Fig. 2f, g). Together, these results indicate that the failure of oncogenic RAS to induce macropinocytosis in V-ATPase-depleted cells can be attributed to impaired plasma membrane localization of RAC1 owing to perturbed cholesterol trafficking. Consistent with this conclusion, several screen hits correspond to components of regulatory pathways controlling cholesterol metabolism and trafficking, including peroxisome proliferator-activated receptor subunits (PPAR α , PPAR γ and PPAR δ)¹⁰ and StAR-related lipid transfer protein 4¹¹.

RAS regulates V-ATPase localization

We next investigated the effect of oncogenic RAS on the subcellular distribution of V-ATPase. Whereas in HeLa cells harbouring wild-type RAS, localization of V-ATPase was predominantly cytoplasmic, expression of

HRAS(G12V) or KRAS(G12V) in these cells led to pronounced accumulation of V-ATPase at the plasma membrane, as determined by immunofluorescence using a V1A subunit-specific antibody (Fig. 2a, b, Extended Data Fig. 3a, b), and substantiated by subcellular fractionation (Fig. 2c). Furthermore, depletion of mutant KRAS by siRNA in lung, pancreatic and colon cancer cells was accompanied by loss of plasma membrane-associated V-ATPase, indicating an essential role for mutant RAS in maintaining plasma membrane pools of V-ATPase (Fig. 2d). In line with these observations, immunohistochemical staining of V-ATPase in human pancreatic ductal adenocarcinoma (PDAC) specimens revealed prominent staining at the cell periphery in neoplastic lesions, in contrast to the predominantly cytoplasmic staining observed in adjacent normal tissues (Fig. 2e). Thus, mutant RAS-dependent plasma membrane V-ATPase displayed preferential accumulation in membrane ruffles, consistent with patterns observed in invasive breast, melanoma and pancreatic cancer cells^{4,12–14}.

Mammalian V-ATPase is composed of two domains. The peripheral V1 domain is composed of eight subunits (A–H) and is responsible for ATP hydrolysis, whereas the membrane-embedded V0 domain is comprised of six subunits (a, c, c', d and e) and is responsible for proton translocation⁴. Targeting of V-ATPase to different cellular membranes is controlled by isoforms of subunit a (V0a1–V0a4), with V0a3 and V0a4 being responsible for directing V-ATPase complexes to the plasma membrane⁴. We established a role for V0a3 in mediating mutant RAS-dependent plasma membrane localization on the basis of the observation that expression of HRAS(G12V) or KRAS(G12V) led to an increase in pools of plasma membrane-associated V0a3 (Fig. 2f), and suppression of V0a3 expression via siRNA-mediated targeting was associated with preferential loss of plasma membrane-associated V-ATPase (Fig. 2g–i, Extended Data Fig. 3c). We used the latter approach to further assess the specific role of plasma membrane-associated V-ATPase in RAS-induced macropinocytosis. Silencing of V0a3 resulted in inhibition of macropinocytosis in cells expressing mutant RAS (Fig. 2j, Extended

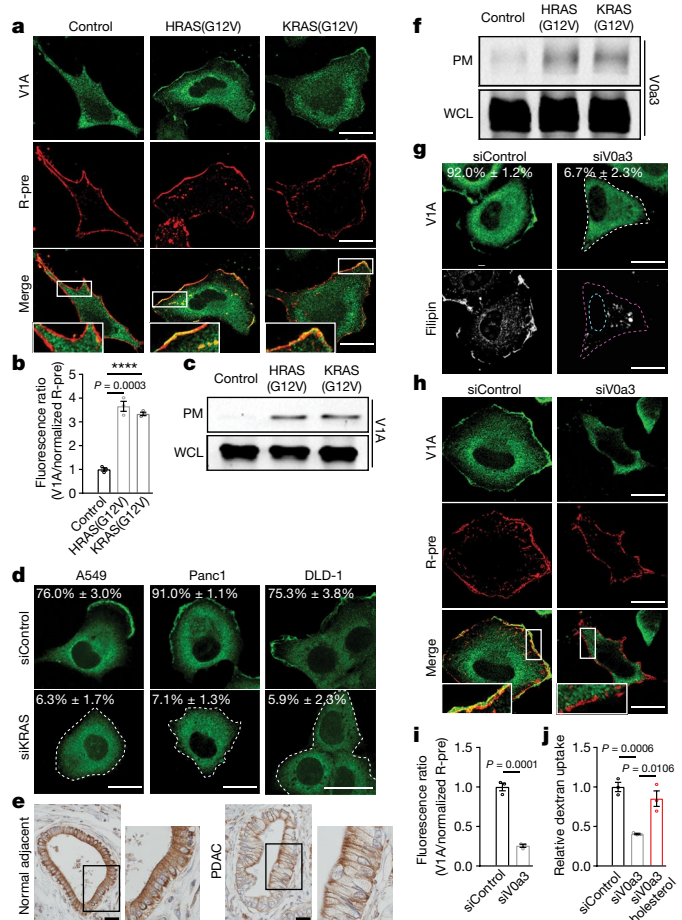


Fig. 2 | Oncogenic RAS induces the translocation of V-ATPase to the plasma membrane. **a–d**, Effect of oncogenic RAS expression on V-ATPase localization. **a**, Fluorescence micrographs of HeLa control, HRAS(G12V) and KRAS(G12V) cells immunostained for V1A subunit; membranes are labelled with R-pre (see Methods). **b**, Quantification of the ratio of V1A to R-pre membrane localization from **a**. **c**, Immunoblots of V1A in the plasma membrane (PM) fraction and whole-cell lysate (WCL) from HeLa control, HRAS(G12V) and KRAS(G12V) cells. **d**, Fluorescence micrographs of siRNA-transfected mutant RAS cells immunostained for V1A. **e**, Immunohistochemical staining of V1A in a section from a PDAC tumour and a section from normal adjacent tissue. Images are representative of staining patterns observed in 11 out of 12 patients. **f**, Immunoblots of V0a3 in the plasma membrane fraction and whole-cell lysate from HeLa control, HRAS(G12V) and KRAS(G12V) cells. **g**, Fluorescence micrographs of V1A immunostaining and Filipin labelling of siRNA-transfected HeLa HRAS(G12V) cells. **h**, Fluorescence micrographs of V1A immunostaining and membrane labelling with R-pre of siRNA-transfected HeLa HRAS(G12V) cells. **i**, Quantification of the ratio of V1A to R-pre membrane localization from **h**. **j**, Quantification of TMR-dextran uptake in HeLa HRAS(G12V) cells transfected with siRNAs and treated as indicated. Images (**a**, **d**, **g**, **h**) and immunoblots (**c**, **f**) are representative of three biological replicates. In **d**, **g**, data are mean \pm s.e.m., representing the fraction of cells that display V1A plasma membrane localization; dashed lines delineate the cell and/or nucleus. Inset regions are enlarged (**a**, **h**, **e**). Scale bars, 10 μ m (**a**, **d**, **g**, **h**) and 50 μ m (**e**). At least 500 (**d**, **g**, **j**) and 7 (**b**, **i**) cells were quantified in each biological replicate ($n = 3$); data are mean \pm s.e.m.; unpaired two-tailed Student's *t*-test; **** $P < 0.0001$. For **c**, **f**, gel source data are provided in Supplementary Fig. 1.

Data Fig. 3d), indicating that oncogenic RAS-dependent recruitment of V-ATPase to the plasma membrane constitutes an essential step in induction of macropinocytosis. Notably, RAS-mutant cells in which plasma membrane localization of V-ATPase was specifically abrogated displayed loss of plasma membrane cholesterol, whereas addition

of cholesterol was sufficient to rescue the defect in macropinocytosis observed in these cells (Fig. 2g, j, Extended Data Fig. 3c, d). Thus, cholesterol trafficking in RAS-mutant cells is critically dependent on plasma membrane-localized V-ATPase.

The abundance of plasma membrane cholesterol is principally regulated by efflux of cholesterol from endosomes, a process requiring endosomal acidification¹⁵. Since it is well established that V-ATPase molecules shuttle between the plasma membrane and endosomes^{16,17}, we investigated the relationship between plasma membrane and endosomal pools of V-ATPase in RAS-mutant cells. Using quantitative colocalization analysis, we found that expression of mutant RAS resulted in significant enrichment of endosomal populations that were positive for both V-ATPase and the endosomal marker Rab7 (Extended Data Fig. 4a–c). This enrichment was substantially reduced when plasma membrane V-ATPase was specifically eliminated by silencing of V0a3 (Extended Data Fig. 4d), suggesting that enrichment of endosomal V-ATPase observed in mutant RAS cells is mediated by endocytic trafficking of plasma membrane-associated V-ATPase. Accordingly, blocking endocytic internalization of plasma membrane V-ATPase using the endocytosis inhibitor dynasore inhibited mutant RAS-dependent accumulation of endosomal V-ATPase (Extended Data Fig. 4b, c). Moreover, using self-complementing split-fluorescent-protein technology to monitor trafficking of V-ATPase in real time, we found that plasma membrane V-ATPase is internalized into intracellular vesicular structures, including Rab7-positive endosomes (Extended Data Fig. 4e, f). Functionally, increased endosomal pools of V-ATPase could promote cholesterol transport by contributing to endosomal acidification¹⁵. Collectively, these results indicate that mutant RAS modulates cholesterol trafficking by affecting partitioning of V-ATPase into subcellular compartments that control cholesterol transport.

RAS–bicarbonate signalling and macropinocytosis

Next, we sought to identify the steps downstream of oncogenic RAS that mediate plasma membrane translocation of V-ATPase. Regulation of plasma membrane accumulation of V-ATPase has been extensively studied in epithelial cells that rely on plasma membrane localization of the pump for extracellular acidification, including renal intercalated cells, in which V-ATPase is responsible for acid secretion into urine, and epididymal clear cells, in which acid secretion facilitates sperm maturation⁴. In these cells, trafficking of V-ATPase to the plasma membrane has been shown to be mediated by activation of bicarbonate-dependent soluble adenylyl cyclase (sAC), which increases cAMP levels, thereby activating protein kinase A (PKA)⁴. To test the contribution of the sAC–PKA pathway to oncogenic RAS-dependent translocation of V-ATPase to the plasma membrane, cells harbouring ectopically expressed or endogenous mutant RAS were treated with the PKA inhibitor H89 and the sAC inhibitor KH7. Both agents specifically inhibited plasma membrane localization of V-ATPase, as determined by microscopic and biochemical analysis (Fig. 3a–e). Disruption of V-ATPase membrane localization was accompanied by loss of plasma membrane cholesterol and reduced macropinocytosis, which could be rescued by addition of exogenous cholesterol (Fig. 3f–i, Extended Data Fig. 5a, b). Additionally, sAC–PKA inhibition led to impaired plasma membrane localization of RAC1 without affecting its activation (Extended Data Fig. 5c, d). Of note, neither KH7 nor H89 had an effect on cholesterol distribution in BxPC-3 human pancreatic cancer cells or HeLa cells harbouring wild-type RAS (Extended Data Fig. 5e–g). However, on introduction of oncogenic RAS, BxPC-3 cells displayed a marked change in cholesterol distribution following KH7 and H89 treatment (Extended Data Fig. 5g). Together, these results indicate that the observed consequences of perturbing the sAC–PKA axis reflect the specific function of this axis downstream of oncogenic RAS and upstream of V-ATPase and cholesterol-dependent macropinocytosis. This conclusion is further supported by our finding that the impaired macropinocytosis and plasma membrane localization of V-ATPase and cholesterol observed in KH7-treated RAS-mutant cells

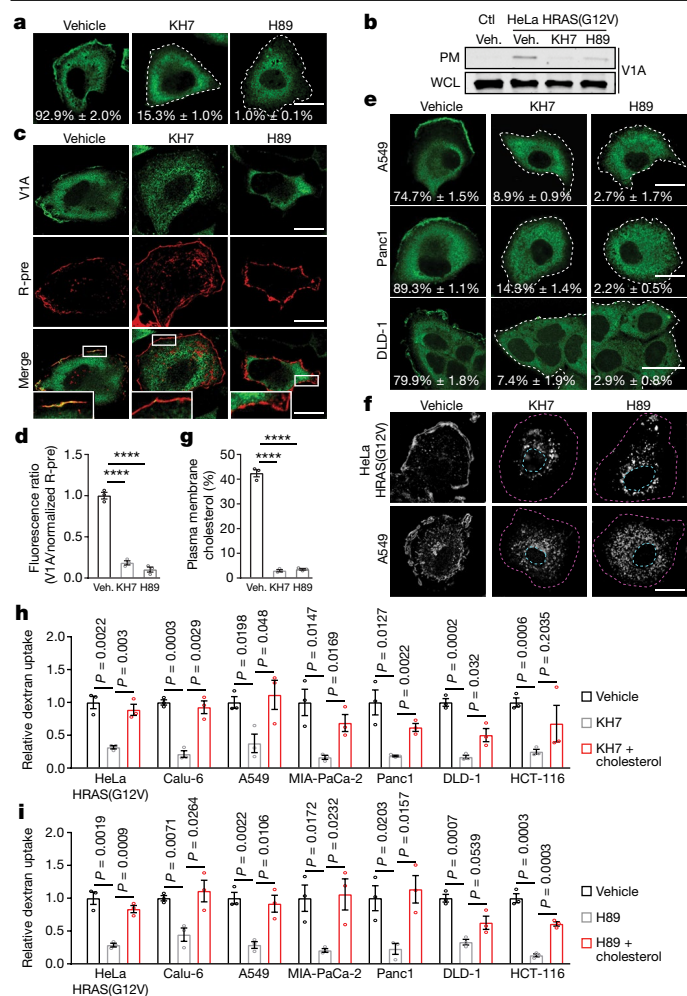


Fig. 3 | Oncogenic RAS-induced macropinocytosis is dependent on sAC-PKA pathway. **a–e**, Effect of inhibition of sAC (with KH7) or PKA (with H89) on V-ATPase membrane translocation. **a, e**, Immunofluorescence staining pattern of V1A in HeLa HRAS(G12V) (**a**) and mutant RAS (**e**) cells treated as indicated. **b**, Immunoblots of V1A expression in the plasma membrane fraction and whole-cell lysate from HeLa control and HRAS(G12V) cells treated as indicated. Gel source data are shown in Supplementary Fig. 1. **c**, Fluorescence micrographs of HeLa HRAS(G12V) cells treated as indicated with V1A immunostaining and membrane labelling with R-pre. Inset region is enlarged. **d**, Quantification of the ratio of V1A to R-pre membrane localization from **c**. Veh., vehicle. **f**, Fluorescence micrographs of cholesterol (filipin) distribution in cells treated as indicated. **g**, Quantification of plasma membrane cholesterol in HeLa HRAS(G12V) cells treated as indicated. **h, i**, Quantification of TMR-dextran uptake in mutant RAS cells following treatment with vehicle, KH7 (**h**) or H89 (**i**) with or without exogenous cholesterol. Images (**a, c, e, f**) and immunoblot (**b**) are representative of three biological replicates. In **a, e**, the data are mean \pm s.e.m., representing the percentage of cells that display V1A plasma membrane localization. Scale bars, 10 μ m. Dashed lines delineate cell and/or nucleus. At least 500 (**a, e, h, i**), 7 (**d**) and 50 (**g**) cells were quantified in each biological replicate ($n=3$); data are mean \pm s.e.m.; unpaired two-tailed Student's *t*-test; **** $P<0.0001$.

could be rescued by expression of constitutively active PKA (CA-PKA) (Extended Data Fig. 6a, b). This rescuing effect was abrogated when cells were depleted of plasma membrane V-ATPase by knockdown of V0a3, indicating that the effects of PKA are mediated by plasma membrane V-ATPase (Extended Data Fig. 6c, d). It is worth noting that CA-PKA failed to rescue macropinocytosis in KRAS-knockdown cells, indicating that whereas PKA activation is necessary for mutant KRAS-dependent plasma membrane translocation of V-ATPase and induction of macropinocytosis, it is not sufficient (Extended Data Fig. 6e).

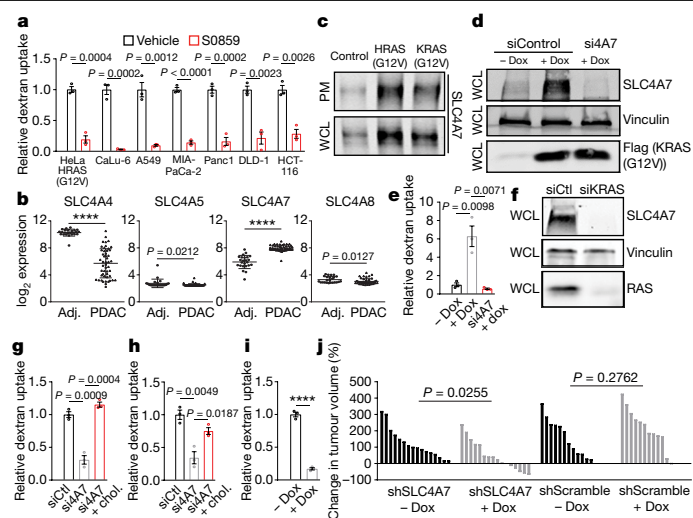


Fig. 4 | SLC4A7 is necessary for RAS-induced macropinocytosis and tumour growth. **a**, Quantification of TMR-dextran uptake following treatment of mutant RAS cells as shown. **b**, mRNA transcript levels of SLC4 family members in PDAC compared with normal adjacent (Adj.) tissue from 74 patients ($n=50$ PDAC, $n=24$ adjacent normal). **c**, Effect of oncogenic RAS expression on SLC4A7 protein levels. Immunoblots of SLC4A7 in the plasma membrane fraction and whole-cell lysate from HeLa control, HRAS(G12V) or KRAS(G12V) cells. **d, e**, Effect of doxycycline-inducible Flag-KRAS(G12V) (+Dox) expression in BxPC-3 cells on SLC4A7 expression and macropinocytosis. **d**, Immunoblot of SLC4A7 expression (**d**; vinculin loading control) and quantification of FITC-dextran uptake (**e**) with or without doxycycline treatment following SLC4A7 knockdown in BxPC-3 cells. **f**, Immunoblot of SLC4A7 expression (vinculin loading control) following KRAS knockdown in MIA-PaCa-2 cells. **g, h**, Quantification of TMR-dextran uptake following SLC4A7 knockdown in HeLa HRAS(G12V) (**g**) and MIA-PaCa-2 (**h**) cells treated as shown. **i, j**, Effect of doxycycline-inducible SLC4A7 depletion in MIA-PaCa2 cells on macropinocytosis and tumour growth. **i**, Quantification of FITC-dextran uptake treated as shown. **j**, Waterfall plots of xenografts treated as shown relative to baseline. Each bar represents a tumour. Immunoblots (**c, d, f**) are representative of three biological replicates. At least 500 (**a, e, g–i**) cells were quantified in each biological replicate ($n=3$); data are mean \pm s.e.m.; unpaired two-tailed Student's *t*-test; **** $P<0.0001$. Gel source data for **c, d, f** are shown in Supplementary Fig. 1.

RAS–bicarbonate signalling and tumorigenesis

Given the exquisite dependence of sAC activation on bicarbonate¹⁸, and the requirement of sAC activity for oncogenic RAS-induced macropinocytosis, we set out to identify the source of bicarbonate that potentially contributes to this process. Incubation of cells harbouring ectopically expressed or endogenous mutant RAS in bicarbonate-free medium resulted in inhibition of dextran uptake, indicating an essential role for extracellular bicarbonate in RAS-induced, sAC–PKA-dependent macropinocytosis (Extended Data Fig. 7a). Transmembrane flux of bicarbonate is facilitated by bicarbonate transport proteins¹⁹; several bicarbonate transporters have been implicated in cancer including the SLC4 family of co-transporters^{19,20}. Treatment of mutant RAS cells with the SLC4-family inhibitor S0859 led to significant inhibition of macropinocytosis (Fig. 4a), and mutant RAS-dependent plasma membrane localization of V-ATPase was abrogated in S0859-treated cells (Extended Data Fig. 7b–d). Expression of CA-PKA in S0859-treated mutant RAS cells restored macropinocytosis and plasma membrane localization of V-ATPase and cholesterol (Extended Data Fig. 7e, f) consistent with the role of SLC4 transporters in regulating macropinocytosis via sAC-mediated activation of PKA. Treatment of BxPC-3 cells with S0859 had no effect on cholesterol distribution (Extended Data Fig. 7g); however, expression of mutant RAS in these cells restored sensitivity to the inhibitor, indicating that the dependency on SLC4 transporters is linked to oncogenic mutant RAS signalling (Extended Data Fig. 7g).

The SLC4 family comprises ten genes, of which five—SLC4A4, SLC4A5, SLC4A7, SLC4A8 and SLC4A10—are Na⁺-coupled bicarbonate transporters that mediate bicarbonate import across the plasma membrane^{19,20}. Analysis of the expression of these transporters in human PDAC datasets revealed that SLC4A7 was uniquely upregulated in PDAC tumours (Fig. 4b). We next investigated the causal relationship between oncogenic RAS and SLC4A7 expression. As illustrated in Fig. 4c–e and Extended Data Fig. 7h, ectopic expression of HRAS(G12V) and KRAS(G12V) in wild-type RAS cells (HeLa and BxPC-3) led to increased SLC4A7 expression at both the mRNA and protein levels, the latter being reflected in both total and plasma membrane protein levels. Conversely, depletion of mutant KRAS expression by siRNA in PDAC cells harbouring mutant RAS resulted in decreased SLC4A7 mRNA and protein (Fig. 4f, Extended Data Fig. 7i). Notably, SLC4A7 depletion by siRNA resulted in inhibition of macropinocytosis, an effect that could be rescued by addition of cholesterol (Fig. 4g, h). This suggests that the dependence of mutant RAS-induced macropinocytosis on SLC4A7 is linked to its essential role in mediating mutant RAS-driven cholesterol trafficking to the plasma membrane. Our interpretation is supported by the observation that the sensitivity of cholesterol transport to inhibition of SLC4-mediated bicarbonate transport is restricted to cells expressing mutant RAS (Extended Data Fig. 7g).

Upregulation of SLC4A7 has been observed in human breast cancer, where it has been linked to ErbB receptor-mediated signalling²¹. Little is known about the relationship between deregulated ErbB signalling and macropinocytosis; however, it is well known that ligand-induced activation of EGFR, a member of the ErbB receptor family, stimulates macropinocytosis²². Furthermore, PI3K–Akt signalling, a critical effector pathway for RAS-induced macropinocytosis²³, and MEK–ERK signalling have been implicated in ErbB-dependent upregulation of SLC4A7²¹. Consistent with these observations, we have found that mutant RAS-dependent upregulation of SLC4A7 was abrogated upon treatment with inhibitors of PI3K and MEK, indicating the importance of both effector pathways in this initial step of mutant RAS-dependent macropinocytosis (Extended Data Fig. 7j).

Maintenance of submembranous alkaline pH (pH_{sm}) has been shown to be essential for the actin cytoskeleton dynamics that regulate membrane ruffling and macropinocytosis²⁴. To test the extent to which the identified roles of V-ATPase and SLC4A7 in RAS-induced macropinocytosis could be attributed to effects on pH_{sm}, we exploited a dual-emission fluorescent construct²⁴ to ratiometrically measure pH_{sm}. Our results demonstrate that depletion of SLC4A7 or V-ATPase does not lead to decreased pH_{sm}, indicating that their requirement for mutant RAS-induced macropinocytosis is not linked to alterations in pH_{sm} homeostasis (Extended Data Fig. 7k).

Collectively, our observations implicate SLC4A7 as an essential mediator of RAS-induced macropinocytosis, a process previously shown to support tumour growth¹. Therefore, we sought to directly evaluate the role of SLC4A7 in tumour progression using a xenograft mouse model¹. A MIA-PaCa-2 stable human pancreatic cancer cell line harbouring a doxycycline-inducible short hairpin RNA (shRNA) targeting SLC4A7, which shows decreased macropinocytosis on knockdown (Fig. 4i, Extended Data Fig. 7l), was implanted into flanks of nude mice. When tumours reached an average volume of 50–100 mm³, doxycycline was administered to induce SLC4A7 knockdown. After 14 days, tumours induced with doxycycline displayed attenuated growth relative to control tumours and, in some cases, regression (Fig. 4j). Notably, doxycycline treatment in control MIA-PaCa-2 cells harbouring scrambled shRNA had no effect on tumour progression. Furthermore, using the same approach with doxycycline-inducible knockdown of SLC4A7 in BxPC-3 cells expressing wild-type RAS had no impact on tumour growth (Extended Data Fig. 7m, n). These results identify an essential role for SLC4A7 in mutant RAS-dependent pancreatic tumour growth.

This study implicates plasma membrane V-ATPase as a key mediator of oncogenic RAS-induced macropinocytosis. Engagement of the bicarbonate-dependent SAC–PKA signalling axis as the principal mechanism by

which oncogenic RAS promotes membrane accumulation of V-ATPase may enable the coupled regulation of pH homeostasis and metabolic adaptation in mutant RAS-driven tumours. Plasma membrane V-ATPases have been implicated in regulation of tumour cell drug sensitivity and invasive capacity⁴. Thus, in addition to its role in macropinocytosis, oncogenic RAS-induced translocation of V-ATPase to the plasma membrane might have a broader impact on critical fitness features of mutant RAS cells. Moreover, selective targeting of plasma membrane V-ATPase could offer a strategy for the development of therapeutics against mutant RAS tumours.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1831-x>.

1. Commisso, C. et al. Macropinocytosis of protein is an amino acid supply route in Ras-transformed cells. *Nature* **497**, 633–637 (2013).
2. Kamphorst, J. J. et al. Human pancreatic cancer tumors are nutrient poor and tumor cells actively scavenge extracellular protein. *Cancer Res.* **75**, 544–553 (2015).
3. Fennell, M., Commisso, C., Ramirez, C., Garippa, R. & Bar-Sagi, D. High-content, full genome siRNA screen for regulators of oncogenic HRAS-driven macropinocytosis. *Assay Drug Dev. Technol.* **13**, 347–355 (2015).
4. Stransky, L., Cotter, K. & Forgac, M. The function of v-ATPases in cancer. *Physiol. Rev.* **96**, 1071–1091 (2016).
5. Furuchi, T., Aikawa, K., Arai, H. & Inoue, K. Bafilomycin A₁, a specific inhibitor of vacuolar-type H⁺-ATPase, blocks lysosomal cholesterol trafficking in macrophages. *J. Biol. Chem.* **268**, 27345–27348 (1993).
6. Iliev, A. I., Djannatian, J. R., Nau, R., Mitchell, T. J. & Wouters, F. S. Cholesterol-dependent actin remodeling via RhoA and Rac1 activation by the *Streptococcus pneumoniae* toxin pneumolysin. *Proc. Natl Acad. Sci. USA* **104**, 2897–2902 (2007).
7. del Pozo, M. A. et al. Integrins regulate Rac targeting by internalization of membrane domains. *Science* **303**, 839–842 (2004).
8. Kozik, P. et al. A human genome-wide screen for regulators of clathrin-coated vesicle formation reveals an unexpected role for the V-ATPase. *Nat. Cell Biol.* **15**, 50–60 (2013).
9. Plowman, S. J., Muncke, C., Parton, R. G. & Hancock, J. F. H-ras, K-ras, and inner plasma membrane raft proteins operate in nanoclusters with differential dependence on the actin cytoskeleton. *Proc. Natl Acad. Sci. USA* **102**, 15500–15505 (2005).
10. Li, T. & Chiang, J. Y. L. Regulation of bile acid and cholesterol metabolism by PPARs. *PPAR Res.* **2009**, 501739 (2009).
11. Garbarino, J. et al. STARD4 knockdown in HepG2 cells disrupts cholesterol trafficking associated with the plasma membrane, ER, and ERC. *J. Lipid Res.* **53**, 2716–2725 (2012).
12. Capecchi, J. & Forgac, M. The function of vacuolar ATPase (V-ATPase) a subunit isoforms in invasiveness of MCF10a and MCF10CA1a human breast cancer cells. *J. Biol. Chem.* **288**, 32731–32741 (2013).
13. Nishishio, T. et al. The a3 isoform vacuolar type H⁺-ATPase promotes distant metastasis in the mouse B16 melanoma cells. *Cancer Res.* **9**, 845–855 (2011).
14. Chung, C. et al. The vacuolar-ATPase modulates matrix metalloproteinase isoforms in human pancreatic cancer. *Lab. Invest.* **91**, 732–743 (2011).
15. Deffieu, M. S. & Pfeffer, S. R. Niemann–Pick type C1 function requires luminal domain residues that mediate cholesterol-dependent NPC2 binding. *Proc. Natl Acad. Sci. USA* **108**, 18932–18936 (2011).
16. Breton, S., Lisanti, M. P., Tyszkowski, R., McLaughlin, M. & Brown, D. Basolateral distribution of caveolin-1 in the kidney. Absence from H⁺-ATPase-coated endocytic vesicles in intercalated cells. *J. Histochem. Cytochem.* **46**, 205–214 (1998).
17. Breton, S. & Brown, D. Regulation of luminal acidification by the V-ATPase. *Physiology (Bethesda)* **28**, 318–329 (2013).
18. Chen, Y. et al. Soluble adenylate cyclase as an evolutionarily conserved bicarbonate sensor. *Science* **289**, 625–628 (2000).
19. Gorbatenko, A., Olesen, C. W., Boedtkjer, E. & Pedersen, S. F. Regulation and roles of bicarbonate transporters in cancer. *Front. Physiol.* **5**, 130 (2014).
20. Romero, M. F., Chen, A.-P., Parker, M. D. & Boron, W. F. The SLC4 family of bicarbonate (HCO₃⁻) transporters. *Mol. Aspects Med.* **34**, 159–182 (2013).
21. Gorbatenko, A. et al. ErbB2 upregulates the Na⁺/HCO₃⁻-cotransporter NBCn1/SLC4A7 in human breast cancer cells via Akt, ERK, Src, and Kruppel-like factor 4. *FASEB J.* **28**, 350–363 (2014).
22. Haigler, H. T., McKanna, J. A. & Cohen, S. Rapid stimulation of pinocytosis in human carcinoma cells A-431 by epidermal growth factor. *J. Cell Biol.* **83**, 82–90 (1979).
23. Amyere, M. et al. Constitutive macropinocytosis in oncogene-transformed fibroblasts depends on sequential permanent activation of phosphoinositide 3-kinase and phospholipase C. *Mol. Biol. Cell* **11**, 3453–3467 (2000).
24. Koivusalo, M. et al. Amiloride inhibits macropinocytosis by lowering submembranous pH and preventing Rac1 and Cdc42 signaling. *J. Cell Biol.* **188**, 547–563 (2010).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Article

Methods

Reagents and constructs

TMR–Dextran and FITC–Dextran (70 kDa) were purchased from Fina Biosolutions. All chemicals were molecular biology grade. Bafilomycin A1, Dynasore, 5-(*N*-ethyl-*N*-isopropyl)-amiloride (EIPA), LY294002, U0126, KH7, H89 and S0859 were purchased from Cayman Chemical. Cholesterol-water soluble reagent was purchased from Sigma (cholesterol-methyl- β -cyclodextran). Non-targeting siRNA pool no. 1 (D-001206-13), ATP6V1A 3' UTR-targeting siRNA (sense: 5'-GCA AUG GUU UGU UGA GAU AUU-3'), siV1A (MU-017590-00), siV0a3 (M-012198-01), siKRAS (M-005069-00), siRac1 (M-003560-06), siSLC4A7 (M-007586-01), siV1B2 (M-011589-01), siV0a1 (M-017618-01), siV1E1 (M-011590-01) and siV0c (M-017620-02) were purchased from GE Dharmacon. A complete list of siRNA sequences can be found in Supplementary Table 2.

pCGT was used as a mammalian expression vector to express T7-HRAS(G12V) and T7-KRAS(G12V). pTRIPZ lentiviral inducible shRNA plasmid was from GE Dharmacon. peGFP–RAC1(WT) was generated as previously described²⁵. GFP–RAC1(L61), GFP–RAC1(L61) K-tail and R-pre constructs were provided by M. Philips²⁶. The R-pre construct contains a modified sequence of the membrane-targeting domain of KRAS linked to red fluorescent protein.

V1A–Flag construct was cloned from HeLa cDNA into pCMV-3tag-8 vector backbone using the following primers: fw 5'-GGAGGACTCGAG ACCAGTATGGATTTTCCAAGCTACCC-3', rv 5'-ACCACCGGATCCTT CAAATCTTCAAGGCTACGGAATGC-3'. V1A–GFP11 construct was created by PCR amplifying GFP11 from peGFP–GFP11-clathrin light chain. peGFP–GFP11–Clathrin light chain was a gift from B. Huang (Addgene plasmid 70217). The following primers were used: fw 5'-TCATCAGCGG CCGCGTCGCCACCATGTCTGGGAGGTT-3', rv 5'-ATTAATGCGGCCGCCT ATCCGGATCCGCTGTAATCCCAGC-3'. The PCR-amplified GFP11 was then cloned into the V1A–Flag construct. The reverse primer introduced a stop codon before the C-terminal Flag-tag. The first NotI site was removed and GFP11 was placed into frame with V1A by site-directed mutagenesis using the following primers: fw 5'-CGTAGCCTTGAAGA TGCCGCGCGCTGCCACC-3', rv 5'-GGTGCGACGCGGCCGCGCATCTT CAAGGCTACG-3'.

Lyn–GFP1-10 construct was created by cloning a Lyn-tail sequence into pcDNA3.1–GFP(1-10). pcDNA3.1–GFP(1-10) was a gift from B. Huang (Addgene plasmid 70219). The following oligonucleotides were used: fw 5'-AGCTTGCCACCATGGGATGTATTAAATCAAAAAGGAAAGACGGG ACAG-3', rv 5'-AATTCTGTCCCGTCTTTCCTTTTGATTTAATACATCCCA TGGTGGCA-3'.

Lyn-tailed mCherry–SEpHluorin was a gift from S. Grinstein (Addgene plasmid 32002).

Constitutively active PKA was created by mutating His87 to Gln in a GFP-tagged PKA catalytic subunit construct using site-directed mutagenesis²⁷ according to the manufacturer's instructions (Quick-change, Agilent Technologies).

pTRIPZ Flag–KRAS(G12V) was constructed by inserting a human codon optimized Flag-tagged KRAS(G12V) into the unique AgeI, MluI sites of pTRIPZ, which simultaneously removed RFP and the shRNA targeting region.

Cell culture and transfection

Human cancer cell lines HeLa, CaLu-6, A549, MIA–PaCa-2, Panc-1, DLD-1, HCT-116 and BxPC-3 were obtained and originally authenticated by short tandem repeat (STR) from the American Type Culture Collection. Cell lines are routinely authenticated in-house by cell morphology. All cell lines used tested negative routinely for mycoplasma contamination by DAPI staining. The BxPC-3 stable cell line with doxycycline-inducible expression of Flag–KRAS(G12V) was generated with lentiviral particles in accordance with standard protocols. Cells were transduced with lentiviral particles containing pTRIPZ Flag–KRAS(G12V) and selected with

puromycin (2 μ g ml⁻¹) for three days. All cells were maintained under 5% CO₂ at 37 °C in either DMEM (HeLa, CaLu-6, A549, MIA–PaCa-2, Panc-1, BxPC-3; Invitrogen), RPMI (DLD-1; Invitrogen) or McCoy's (HCT-116; Invitrogen) medium supplemented with 10% FBS (Gibco). Transient transfections were performed by using X-tremeGENE 9 reagent (Sigma), following the manufacturer's recommended protocol. Analyses were performed 48 h after transfection of V1A and RAS cDNA expression constructs and 24 h after transfection of RAC1, PKA and R-pre cDNA expression constructs. Transfection of siRNA was performed by using Lipofectamine RNAiMAX reagent (Thermo Fisher Scientific), following the manufacturer's recommended protocol. Analyses were performed three days after siRNA treatment.

Experimental treatment conditions

Treatment conditions were maintained the same throughout each experimental assay, as follows: KH7 was used at 25 μ M for a 60-min pre-incubation period. H89 was used at 15 μ M for a 60-min pre-incubation period. S0859 was used at 50 μ M for a 90-min pre-incubation period. Cholesterol was used at 10 μ M and co-incubated with drug treatment or added for 90 min before fixation or dextran uptake in siRNA-treated cells. For bicarbonate withdrawal experiments, cells were washed in bicarbonate-free DMEM (Sigma, HEPES-buffered) and then placed in 0 mM or 44 mM bicarbonate conditions for a 90-min pre-incubation period. Dynasore was used at 80 μ M for a 10-min pre-incubation period. Bafilomycin A1 was used at 150 nM for a 60-min pre-incubation period. LY294002 and U0126 were used for 36 h at 25 μ M and 10 μ M, respectively, starting 6 h after transfection of T7-HRAS(G12V) or T7-KRAS(G12V). EIPA was used at 50 μ M for a 30-min pre-incubation period.

Identification of functional clusters

The STRING (v.9.0)²⁸ database was used to identify known and predicted protein–protein interactions between the hits. Only interactions with scores above 0.7 are reported here.

Macropinosome visualization and quantification

Macropinocytosis assays were performed as previously described²⁹. Images were captured using an Axiovert 200 inverted fluorescent microscope (Zeiss). Image analysis and quantification was performed as previously described²⁹.

Immunofluorescence

Cells were seeded onto glass coverslips. Forty-eight to 72 h after cell seeding, cells were serum-starved for 3 h. After serum starvation, cells were fixed with 3.7% formaldehyde for 30 min at room temperature. The following sequential steps were done at room temperature: cells were washed twice with PBS, quenched (50 mM NH₄Cl in PBS) for 10 min, permeabilized (0.1% saponin in PBS) for 10 min, blocked (5% goat serum in PBS) for 30 min. Primary antibody for V1A (Abnova) was used at 1:250 dilution. Primary antibody for Rab7 (Cell Signaling) was used at 1:100 dilution. Secondary antibody was used at 1:1,000 dilution. Cells were DAPI-treated to stain nuclei and coverslips mounted onto slides using DAKO Mounting Media (Agilent). Images were captured using LSM510 META Confocal Microscope (Zeiss). For presentation of microscopy images in the figures, raw images were imported into ImageJ for brightness and contrast enhancements. Cell outlines were delineated by setting the auto threshold (triangle) function on ImageJ and manually drawing the cell periphery.

Colocalization analysis

Images for Rab7 and V1A colocalization were acquired as described above to ensure that no pixel saturation was observed (Extended Data Fig. 4a). The same threshold was set for each image across the entire experiment. The ImageJ plugin, JACoP v.2.0³⁰, was used to calculate the Mander's overlap colocalization coefficient of the V-ATPase with

Rab7 for each cell analysed. In brief, the raw images were denoised using an ImageJ function (despeckle). Background subtraction was performed for quantification of signal within endosomes³¹. Pixel size for background subtraction was determined by Rab7 channel. Segmentation was performed to isolate single cells. Thresholding was set to identify V-ATPase pixel intensity within Rab7-positive endosomes on a per-cell basis. To control for random colocalization, we employed Costes' randomization method. The *P* values in Extended Data Fig. 4c, d were >95%, suggesting that colocalization was highly probable. Representative images presented in Extended Data Fig. 4b were imported into Adobe Photoshop CS and whole-image adjustment of brightness was done using the curves function to display colocalization.

Split GFP complementation assay

Soluble GFP (1–10)³² was targeted to the plasma membrane by the addition of a 16 amino acid Lyn tail³³ to the N terminus. Lyn–GFP(1–10) and V1A–GFP11 were transfected into HeLa HRAS(G12V) cells. The split fluorescent protein only forms a functional GFP fluorophore if V1A–GFP11 is in the plasma membrane. Lyn–GFP(1–10) or V1A–GFP11 transfection alone does not result in fluorescence. Twenty-four hours after transfection, live-cell epifluorescent images were captured using the 488-nm laser of a LSM510 META Confocal Microscope (Zeiss). Z-stack time series were captured and processed in ImageJ.

Submembraneous pH determination

Determination of submembraneous pH (pH_{sm}) was performed as described²⁴. In brief, cells were plated on glass-bottom dishes and transfected with the membrane-targeted SEpHluorin–mCherry construct. Measurements were acquired of the SEpHluorin/mCherry fluorescence emission ratio at the plasma membrane by confocal microscopy in live cells. Calibration was performed with K^+ nigericin buffer as described previously²⁴.

PAK-binding assay

Cells were treated as indicated and the PAK-binding assay was performed as previously described³⁴.

Plasma membrane fractionation

Indicated cell lines were grown to 90% confluency. The plasma membrane protein extraction kit (BioVision) was used to separate the plasma membrane fraction from other cellular membranes according to the manufacturer's recommended protocol.

Immunoblot analysis

Lysates were resolved by SDS–PAGE and transferred to nitrocellulose membranes. Membranes were incubated with vinculin (Sigma, 1:10,000), ATP6V1A (Abnova, 1:1,000), ATP6V0a3 (Novus; 1:1,000), Flag (Sigma, 1:2,000), GFP (Cell Signaling, 1:1,000), T7 (Novagen, 1:10,000), tubulin (Sigma, 1:10,000), RAC1 (BD Transduction Laboratories, 1:1,000), AKT (Cell Signaling, 1:1,000), p-AKT (S473) (Cell Signaling, 1:500), ERK2 (EMD Millipore, 1:2,000), p-ERK1/2 (Cell Signaling, 1:1,000), SLC4A7 (Santa Cruz, 1:200) or KRAS (Santa Cruz, 1:500) primary antibodies followed by Alexa Fluor 680 goat anti-mouse IgG (Life Technologies, 1:10,000) or IRDye 800CW goat anti-rabbit IgG (Li-Cor, 1:10,000) secondary antibodies. Blots were analysed using an Odyssey Classic imager (Li-Cor).

Human pancreas specimens

Samples consisted of 5- μm sections that were cut from formalin-fixed, paraffin-embedded (FFPE) blocks provided by the Center for Biospecimen Research and Development of the New York University Langone Medical Center. All samples were anonymized before being transferred to the investigator's laboratory and therefore met exempt human subject research criteria.

FFPE immunohistochemistry

Immunohistochemistry was performed as previously described³⁵. Primary antibody for V1A was used at 1:150. Slides were examined on a Nikon Eclipse 80i microscope.

Quantitative PCR with reverse transcription

Extraction and reverse transcription of total RNA from cell lines was performed using RNeasy mini kit (QIAGEN) and QuantiTect reverse transcription kit (QIAGEN), respectively. SYBR Green PCR Master Mix (Thermo Fisher Scientific) was used for amplification, and the samples were amplified by a two-step PCR with reverse transcription (RT–qPCR) method and analysed on a Stratagene Mx 3005P using $\Delta\Delta C_t$ analysis. Expression levels were normalized by RPL19.

Primers used for RT–qPCR were: SLC4A7-fw: 5'-GCAAGAAACATTCTGAC CCTCA-3', SLC4A7-rv: 5'-GCTTCCACCATTCCATTACCT-3'; KRAS-fw: 5'-AAGTGTGATTGCTTCTAG-3', KRAS-rv: 5'-ATGTTTTCGAA TTTCTCGGACT-3'; RPL19-fw: 5'-GAATGCCAGAGAAGGTCACA-3' and RPL19-rv: 5'-GCTGTGATACATGTGGCGAT-3'.

Human data generation

One-hundred and thirty human PDAC tumour ($n = 75$) and normal adjacent pancreatic tissue ($n = 55$) mRNA expression profiles generated on the same array (Affymetrix GeneChip Human Genome U133 Plus 2.0) were downloaded from Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/>) (GSE15471³⁶ and GSE16515³⁷). Adjacent normal samples clustering with PDAC tumours and PDAC tumour profiles clustering with adjacent normal samples and duplicates were discarded (as previously described³⁶), for a remainder of $n = 74$ tissues ($n = 50$ PDAC tumour and $n = 24$ normal tissues). Raw data were processed and normalized in one batch using a GC-content background correction robust multi-array average (RMA) algorithm (GC-RMA), performed in R. SLC plots and unpaired Student's *t*-test *P* values were generated in GraphPad Prism (GraphPad Software).

Generation of inducible shSLC4A7 cell lines

Lentiviral particles were generated in accordance with standard protocols. For knockdown experiments, cells were transduced with lentiviral particles containing pTRIPZ scramble shRNA or SLC4A7 shRNA and selected with puromycin ($2\ \mu\text{g ml}^{-1}$) for three days.

Mouse experiments

All animal work was approved by New York University Langone Medical Center Institutional Animal Care and Use Committee (IACUC). For xenograft studies, 2×10^6 MIA–PaCa-2 or BxPC-3 cells stable for pTRIPZ–scramble shRNA or pTRIPZ–SLC4A7 shRNA (1:1 in Matrigel, BD Biosciences) were subcutaneously implanted in both flanks of seven-week-old female athymic nude mice (NCRNU, Taconic). When tumour size reached 50–100 mm^3 , mice were separated into two groups by initial tumour volume (baseline) to allow for similar ranges in initial tumour volume. Investigators were blinded once the mice were separated into experimental and control arms by the mice being given a coded number. During the experiment, one investigator measured the tumour volume and read the coded number to the second investigator who recorded the data for analysis. Mice were given normal or doxycycline feed (1g per kg (body weight)), and feed was replaced every two days. Tumour volume was determined using electronic calipers to measure length (l), width (w), using the formula ($w^2 \times l/2$). Tumour volume was measured twice a week. IACUC criteria for maximum allowable tumour size was approximately 1,500 mm^3 or 1.5 cm in diameter (equivalent to 5% of the body weight of a 25 g mouse), which was never exceeded in these experiments. On the basis of previous experiments using human pancreatic cell lines and xenograft model systems that were used in this study, the sample size was deemed sufficiently powered to detect a statistically significant and biologically relevant effect.

Filipin staining

Cells were seeded and fixed as per the immunofluorescence protocol. Cells were then washed in PBS and stained with filipin (500 $\mu\text{g ml}^{-1}$ in PBS) for 30 min. Filipin solution was washed off cells three times with PBS before mounting on coverslips with DAKO mounting media. Images were acquired on a Zeiss Axiovert 200M. Images within an experiment were taken at the same exposure.

Quantification of plasma membrane cholesterol, V-ATPase and RAC1 localization

Analysis was performed on a per-cell basis. For presentation of microscopy images in figures, raw images were imported into ImageJ for brightness and contrast enhancements. ImageJ was used to quantify fluorescence intensities from raw images. For quantification of plasma membrane V-ATPase and cholesterol, we used a plasma membrane-targeting RFP construct and a quantification approach adapted from a previously described method²⁶. In brief, regions of interest (ROIs) on the plasma membrane and in the cytosol were traced and mean fluorescence intensities for each fluorophore were quantified. The plasma membrane was defined as the outermost region of the cell in which R-pre fluorescent signal was observed. To correct for variations in expression levels of R-pre, the fluorescence intensity for R-pre was calculated according to the formula R-pre plasma membrane/R-pre cytosol (normalized R-pre). The fluorescence ratio of plasma membrane V-ATPase to R-pre was defined as plasma membrane V-ATPase/(normalized R-pre). A value of 1 indicates similar plasma membrane signal intensity and by extension maximal plasma membrane abundance. A value of 0 represents a complete absence in the plasma membrane.

For quantification of plasma membrane cholesterol and GFP-RAC1 without the R-pre marker, the following approaches were used. Images with free cholesterol staining were processed with the Auto Threshold (triangle) function in ImageJ to delineate the cell outline. On the raw images, ROIs were drawn around the entire cell (whole cell, WC cholesterol) and immediately adjacent to the plasma membrane (cytosol cholesterol). Mean fluorescence intensities were determined within each ROI and percent plasma membrane cholesterol was calculated according to the formula $100 \times (\text{WC cholesterol} - \text{cytosol cholesterol}) / (\text{WC cholesterol})$. For images with GFP-RAC1, ROIs were drawn around the plasma membrane (plasma membrane GFP-RAC1) and immediately adjacent to the plasma membrane (cytosol GFP-RAC1). The plasma membrane was defined as the outermost region of the cell in which fluorescent signal was observed. Mean fluorescence intensities were determined at each ROI border and surface GFP-RAC1 was calculated according to the formula $[(\text{plasma membrane GFP-RAC1}) - (\text{cytosol GFP-RAC1})] / (\text{cytosol GFP-RAC1})$. Graphs show surface GFP-RAC1 values relative to control.

Statistical analysis

Unless otherwise indicated, data were analysed using GraphPad Prism built-in tests (unpaired two-tailed Student's *t*-tests). For all graphs, error bars indicate mean \pm s.e.m. for $n \geq 3$ biological replicates and *P* values are shown in the graphs. Numbers of samples analysed per experiment are reported in the respective figure legends.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The full-genome siRNA screen datasets have been deposited at NCBI PubChem under the accession code AID1347130. Uncropped immunoblot images are available in Supplementary Fig. 1. Datasets that support the findings of this study are available in Source Data. Full list of siRNA and shRNA used in this study are in Supplementary Table 1. Full list of primary antibodies used in this study are in Supplementary Table 2.

25. Nimnual, A. S., Taylor, L. J., Nyako, M., Jeng, H. H. & Bar-Sagi, D. Perturbation of cytoskeleton dynamics by the opposing effects of Rac1 and Rac1b. *Small GTPases* **1**, 89–97 (2010).
26. Yeung, T. et al. Receptor activation alters inner surface potential during phagocytosis. *Science* **313**, 347–351 (2006).
27. Orellana, S. A. & McKnight, G. S. Mutations in the catalytic subunit of cAMP-dependent protein kinase result in unregulated biological activity. *Proc. Natl Acad. Sci. USA* **89**, 4726–4730 (1992).
28. Franceschini, A. et al. STRING v9.1: protein–protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808–D815 (2013).
29. Commisso, C., Flinn, R. J. & Bar-Sagi, D. Determining the macropinocytic index of cells through a quantitative image-based assay. *Nat. Protoc.* **9**, 182–192 (2014).
30. Bolte, S. & Cordelières, F. P. A guided tour into subcellular colocalization analysis in light microscopy. *J. Microsc.* **224**, 213–232 (2006).
31. Pike, J. A., Styles, I. B., Rappoport, J. Z. & Heath, J. K. Quantifying receptor trafficking and colocalization with confocal microscopy. *Methods* **115**, 42–54 (2017).
32. Kamiyama, D. et al. Versatile protein tagging in cells with split fluorescent protein. *Nat. Commun.* **7**, 11046 (2016).
33. Sato, I. et al. Differential trafficking of Src, Lyn, Yes and Fyn is specified by the state of palmitoylation in the SH4 domain. *J. Cell Sci.* **122**, 965–975 (2009).
34. Walsh, A. B. & Bar-Sagi, D. Differential activation of the Rac pathway by Ha-Ras and K-Ras. *J. Biol. Chem.* **276**, 15609–15615 (2001).
35. Pylayeva-Gupta, Y., Lee, K. E., Hajdu, C. H., Miller, G. & Bar-Sagi, D. Oncogenic Kras-induced GM-CSF production promotes the development of pancreatic neoplasia. *Cancer Cell* **21**, 836–847 (2012).
36. Badea, L., Herlea, V., Dima, S. O., Dumitrascu, T. & Popescu, I. Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia. *Hepatogastroenterology* **55**, 2016–2027 (2008).
37. Pei, H. et al. FKBP51 affects cancer cell response to chemotherapy by negatively regulating Akt. *Cancer Cell* **16**, 259–266 (2009).

Acknowledgements We thank R. Garippa and M. Fennell (Memorial Sloan Kettering Cancer Center) for their help with RNAi screen data analysis; members of the Bar-Sagi laboratory for their comments and discussions and M. Phillips for sharing cDNA constructs. This work was supported by a grant from The Lustgarten Foundation and National Institutes of Health (NIH)/National Cancer Institute (NCI) (CA210263/CA055360) to D.B.-S. C.R. was supported by a grant from NIH (5 T32 GM007238), A.D.H. was supported by a grant from NIH/NCI (T32CA009161) and E.A.V. was supported by a Canadian Institutes of Health Research Fellowship (146792).

Author contributions C.R., A.D.H. and D.B.-S. conceived the cell biological experiments. C.R. carried out the macropinocytic assays, immunofluorescence and microscopy. A.D.H. carried out the biochemical assays. C.R. and A.D.H. carried out the xenograft experiments. E.A.V. conceived and carried out the human gene expression analysis.

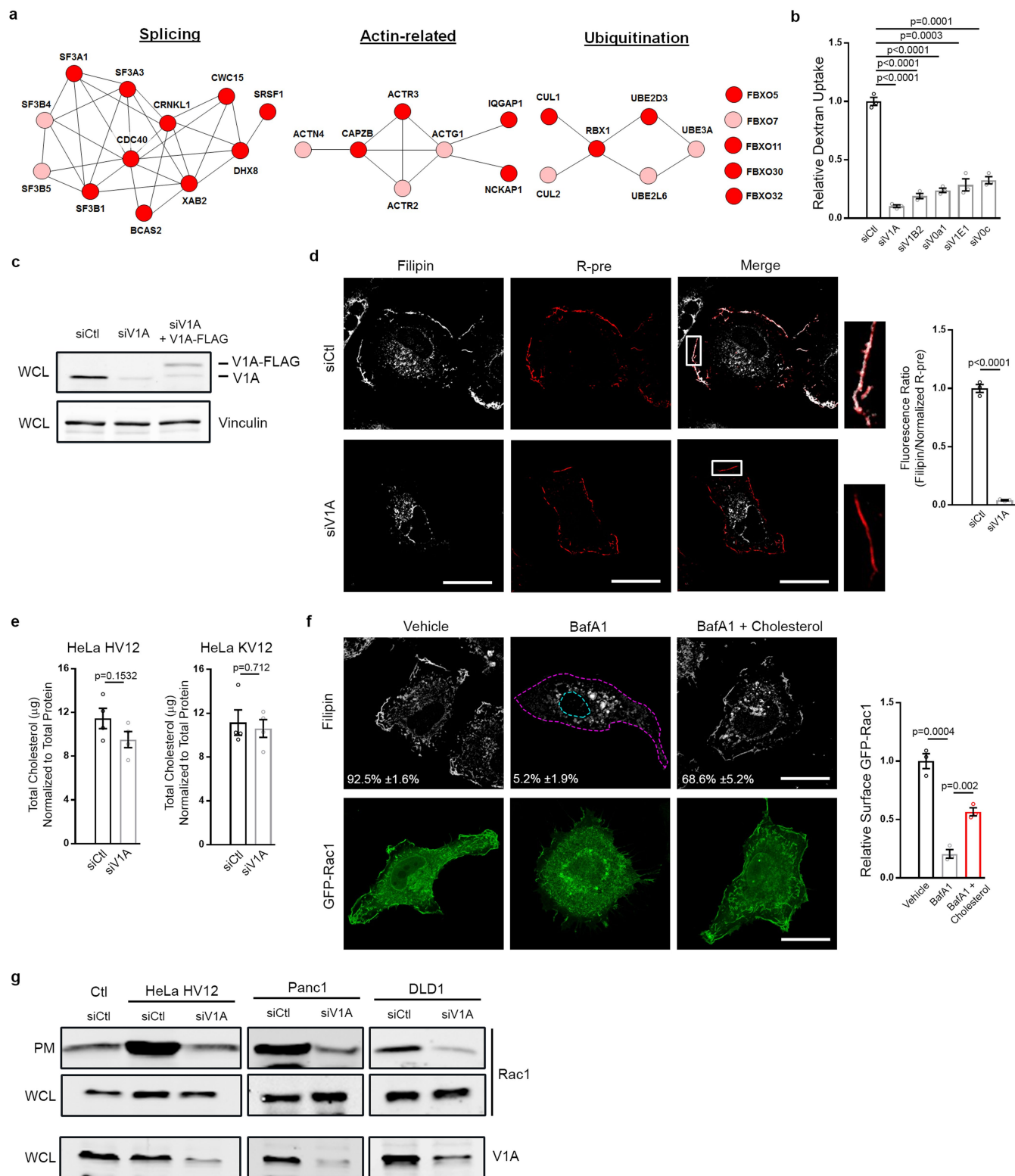
Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1831-x>.

Correspondence and requests for materials should be addressed to D.B.-S.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

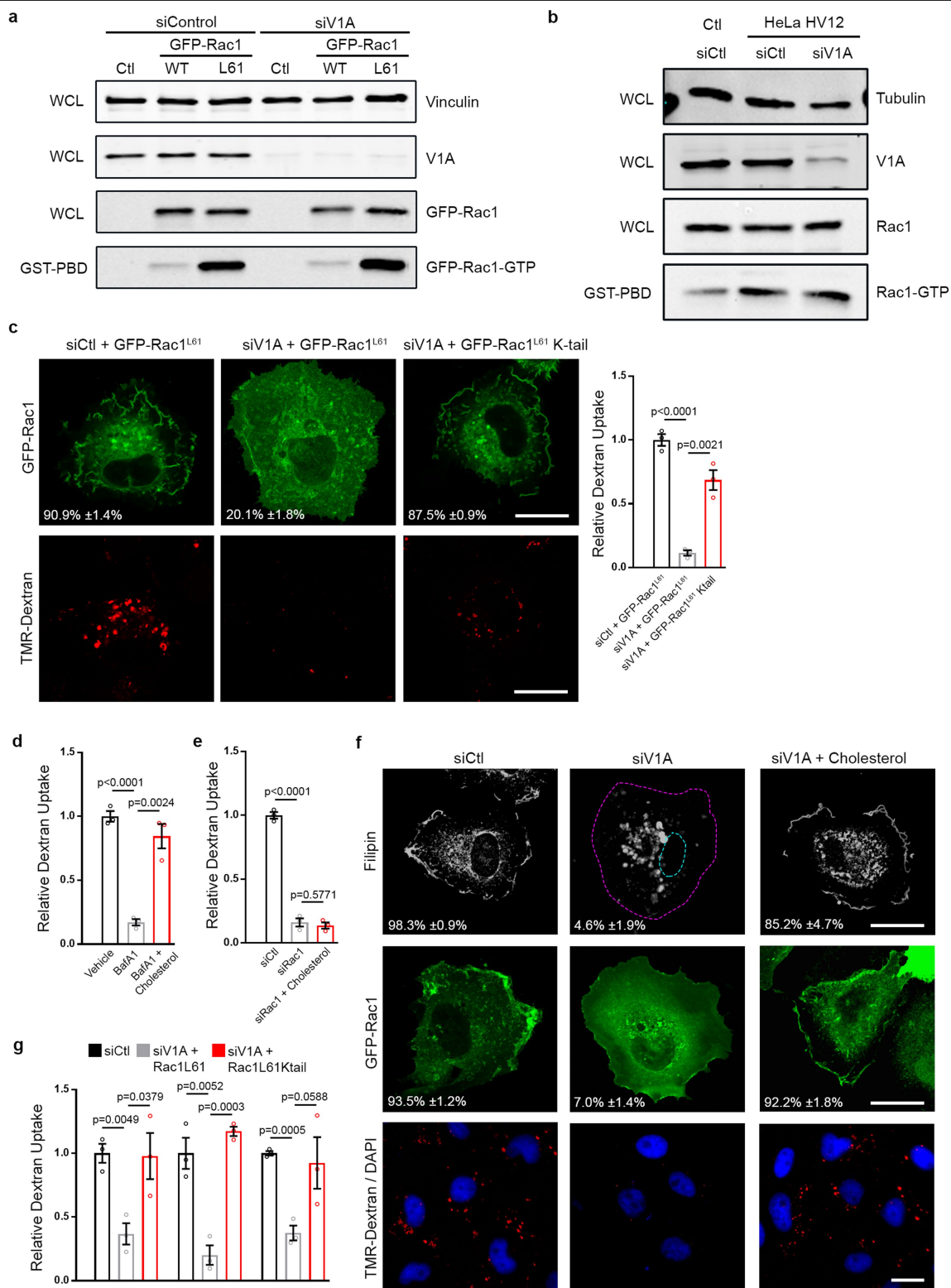


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | V-ATPase is required for RAS-induced

macropinocytosis. **a**, Functional clusters within the macropinocytosis screen hits defined by STRING analysis (pink, primary screen; red, primary and confirmation screen). **b**, Quantification of TMR-dextran uptake following knockdown of the indicated V-ATPase subunits in HeLa HRAS(G12V) cells (HV12). **c**, Immunoblot of V1A expression from whole-cell lysate (vinculin loading control). **d**, Effect of V-ATPase depletion (siV1A) on cholesterol localization in HeLa HRAS(G12V) cells. Fluorescence micrographs of filipin staining (left), membrane labelling with R-pre (a transfected construct containing a modified sequence of the membrane targeting domain of KRAS linked to RFP, middle), merge of filipin and R-pre with boxed areas enlarged to show plasma membrane localization (right), and the quantification of the ratio of filipin to R-pre membrane localization (bar graph) in control (siCtl) or V1A-knockdown conditions. **e**, Effect of V1A knockdown on total cholesterol in HeLa HRAS(G12V) (left) and KRAS(G12V) (KV12) (right) cells. **f**, Effect of bafilomycin

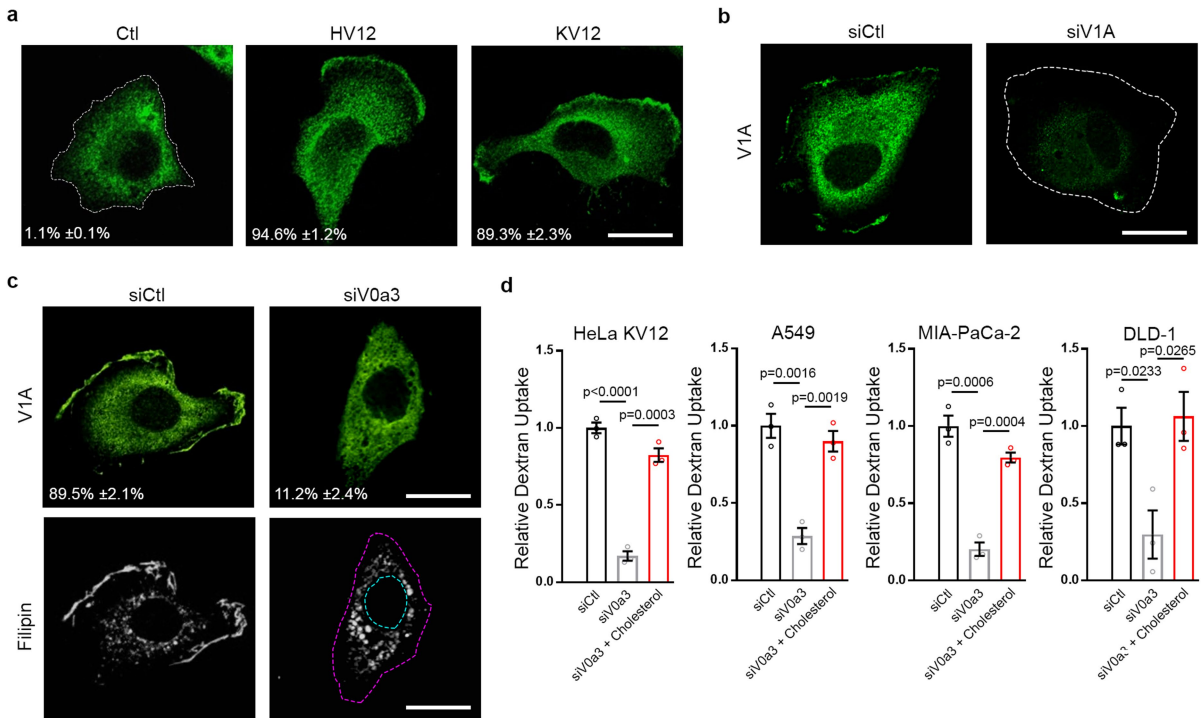
A1 (BafA1) and rescue by exogenous cholesterol on the localization of cholesterol and RAC1 in HeLa HRAS(G12V) cells. Fluorescence micrographs of filipin (top), GFP-RAC1 (bottom) and quantification of relative surface GFP-RAC1 (bar graph). **g**, Effect of oncogenic RAS and V1A expression on RAC1 localization. Immunoblots of RAC1 and V1A in the plasma membrane fraction and whole-cell lysate from HeLa T7-vector control (Ctl) and HRAS(G12V) or oncogenic KRAS cell lines with or without V1A knockdown. Images (**d**, **f**) and immunoblots (**c**, **g**) are representative of three biological replicates. In **f**, the dashed lines delineate the cell and nucleus and data (mean \pm s.e.m.) represent the fraction of cells that display plasma membrane localization of cholesterol. Scale bars, 10 μ m. At least 500 (**b**, **f**) and 20 (**d**) cells were quantified in biological replicates ($n = 3$). In **e**, cholesterol quantification is representative of four biological replicates. All data are mean \pm s.e.m. for the indicated sample size; unpaired two-tailed Student's *t*-test. Gel source data for **c**, **g** are shown in Supplementary Fig. 1.



Extended Data Fig. 2 | See next page for caption.

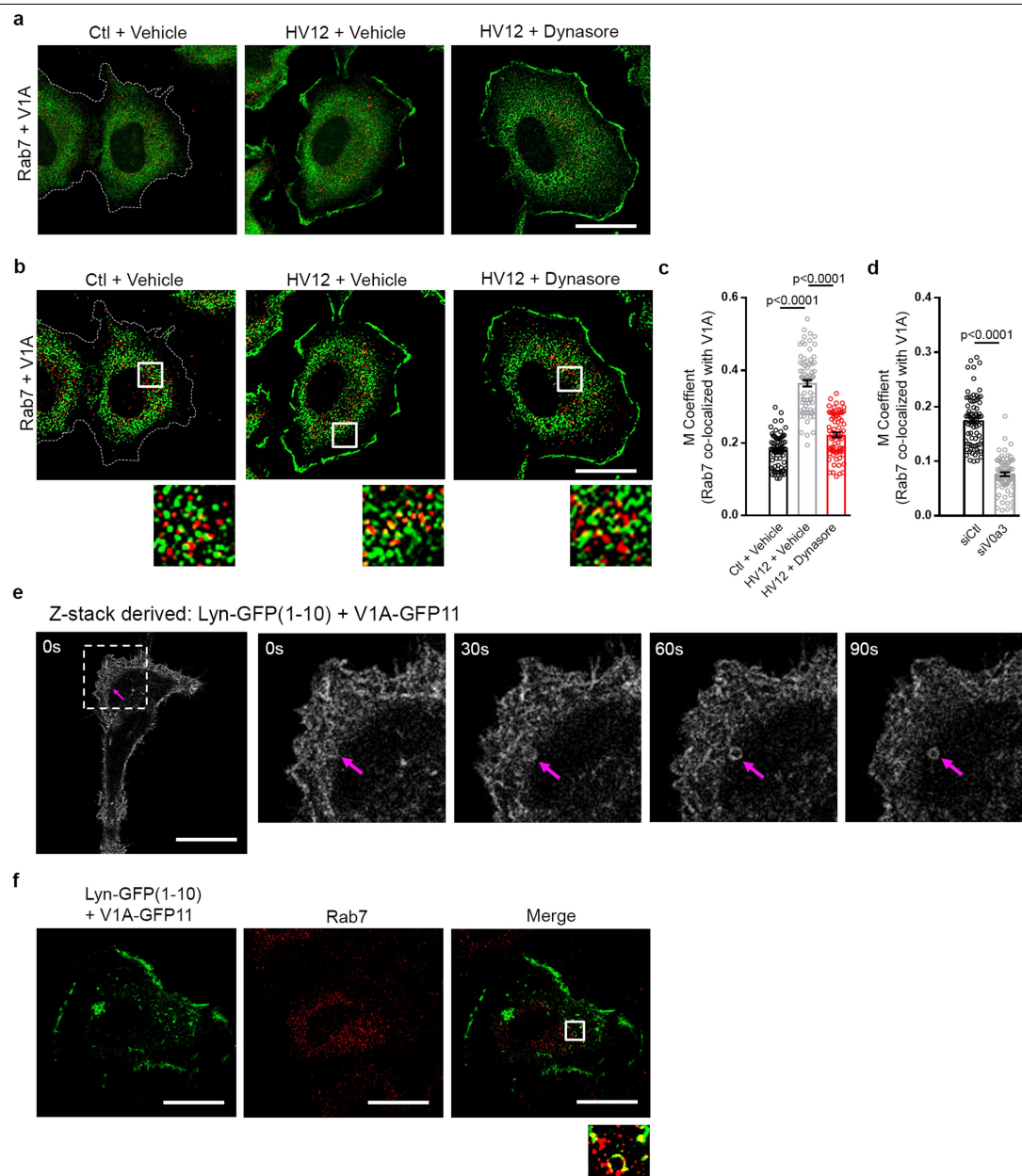
Extended Data Fig. 2 | Plasma membrane-localized RAC1 is required for RAS-induced macropinocytosis. **a, b**, Effect of V1A expression on RAC1 activity. **a**, Immunoblots of RAC1 wild type and RAC1(L61) activity (GST-PBD, pull-down of GFP-RAC1-GTP; vinculin loading control) with or without V1A knockdown in HeLa HRAS(G12V) cells (Ctl, GFP). **b**, Immunoblot of endogenous RAC1 activity (GST-PBD, pull-down of RAC1-GTP; tubulin loading control) in HeLa Ctl and HRAS(G12V) cells with or without V1A knockdown. **c**, Effect of V-ATPase depletion (siV1A) and rescue by plasma membrane-targeted RAC1 (GFP-RAC1(L61) K-tail) on RAC1(L61) localization and macropinocytosis in HeLa HRAS(G12V) cells. Fluorescence micrographs of GFP-RAC1 (top), TMR-dextran uptake (bottom) and quantification of TMR-dextran (bar graph). **d, e**, Effect of V-ATPase or RAC1 depletion on macropinocytosis. Quantification of TMR-dextran uptake following bafilomycin A1 treatment (**d**) or RAC1 knockdown (**e**) in the absence or presence of exogenous cholesterol in HeLa

HRAS(G12V) cells. **f**, Fluorescence micrographs of cholesterol localization (filipin, top), GFP-RAC1 localization (middle) and TMR-dextran uptake (bottom) in V-ATPase-depleted HeLa KRAS(G12V) cells in the absence or presence of exogenous cholesterol. **g**, Quantification of TMR-dextran uptake in mutant RAS cells with V1A knockdown and rescue by plasma membrane-targeted RAC1 (GFP-RAC1(L61) K-tail). Images (**c, f**) and immunoblots (**a, b**) are representative of three biological replicates. Data in **c** are mean \pm s.e.m. representing the fraction of cells that display plasma membrane localization of GFP-RAC1(L61) or GFP-RAC1(L61) K-tail. In **f**, the dashed lines delineate the cell and nucleus and data are mean \pm s.e.m. representing the fraction of cells that display plasma membrane localization of cholesterol (top) or GFP-RAC1 (middle). Scale bars, 10 μ m. At least 500 (**c-g**) cells were quantified in each biological replicate ($n = 3$); data are mean \pm s.e.m.; unpaired two-tailed Student's *t*-test. Gel source data for **a, b** are shown in Supplementary Fig. 1.



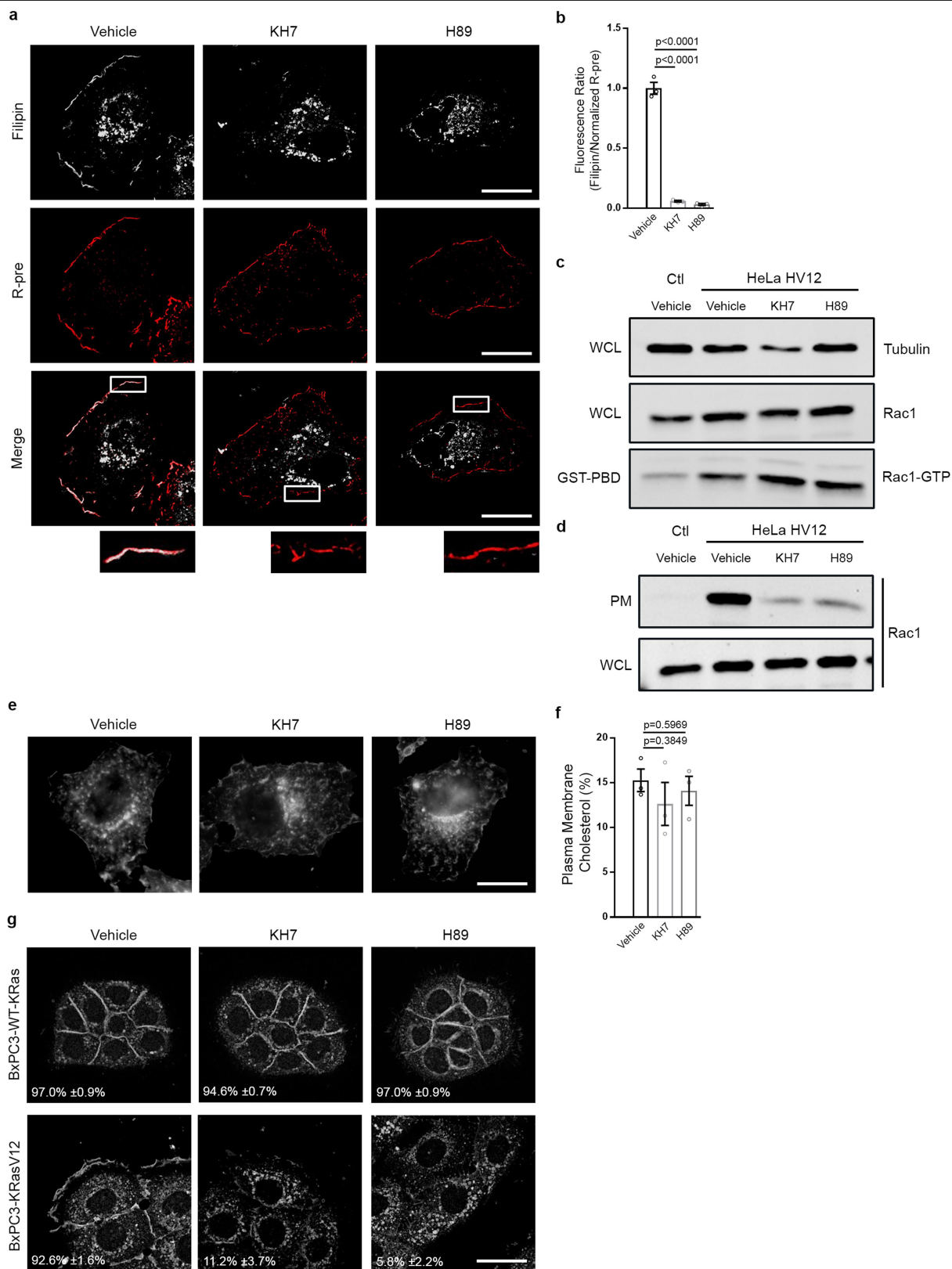
Extended Data Fig. 3 | Plasma membrane V-ATPase regulates cholesterol distribution and macropinocytosis. **a**, Fluorescence micrographs of HeLa control and HeLa HRAS(G12V) or T7-KRAS(G12V) cells immunostained with V1A antibody. **b**, Validation of V1A antibody for immunofluorescence. Fluorescence micrographs of V1A immunostaining of HeLa HRAS(G12V) cells transfected with the indicated siRNAs. **c**, Fluorescence micrographs of V1A immunostaining and filipin labelling of HeLa KRAS(G12V) cells transfected with the indicated siRNAs. **d**, Quantification of TMR-dextran uptake in mutant

RAS cell lines transfected with siV0a3 in the presence or absence of exogenous cholesterol. Images (**a–c**) are representative of three biological replicates. Scale bars, 10 μ m. In **b**, **c**, dashed lines delineate the cell and/or nucleus. Data in **a**, **c**, are mean \pm s.e.m. representing the fraction of cells that display V1A plasma membrane localization. At least 500 (**a**, **c**, **d**) cells were quantified in each biological replicate ($n=3$); data are mean \pm s.e.m.; unpaired two-tailed Student's *t*-test.



Extended Data Fig. 4 | Plasma membrane V-ATPase feeds into Rab7-positive endosomes. **a, b**, Fluorescence micrographs of HeLa control and HRAS(G12V) cells with the indicated treatment showing V1A (green) and Rab7 (red) immunostaining. Representative original image used to calculate Mander's overlap coefficient (**a**) and processed image (**b**). **c, d**, Quantification of Rab7 with V1A colocalization in HeLa cells treated as indicated using Mander's overlap coefficient (M coefficient). **e, f**, Epifluorescence imaging of self-complementing GFP from HeLa HRAS(G12V) cells transfected with V1A-GFP11 and plasma membrane-targeting Lyn-GFP(1-10). Positive fluorescence indicates V1A localization to the plasma membrane. **e**, Fluorescence

micrographs of time-lapse imaging. The boxed area of the cell (left) was enlarged (four right images) to show plasma membrane V-ATPase being internalized over time and forming a vesicle (arrow). Time (in s) is shown on each micrograph. **f**, Fluorescence micrographs of self-complementing GFP (left), immunofluorescence of Rab7 (middle) and merge of V1A and Rab7 (right) with boxed area enlarged beneath the image to show Rab7 colocalization with plasma membrane-derived V1A. In **a, b**, images are representative of three biological replicates. For **c, d**, 25 cells were quantified in each biological replicate ($n = 3$); data are mean \pm s.e.m.; unpaired two-tailed Student's *t*-test. Scale bars, 10 μ m.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | sAC-PKA pathway is necessary for oncogenic

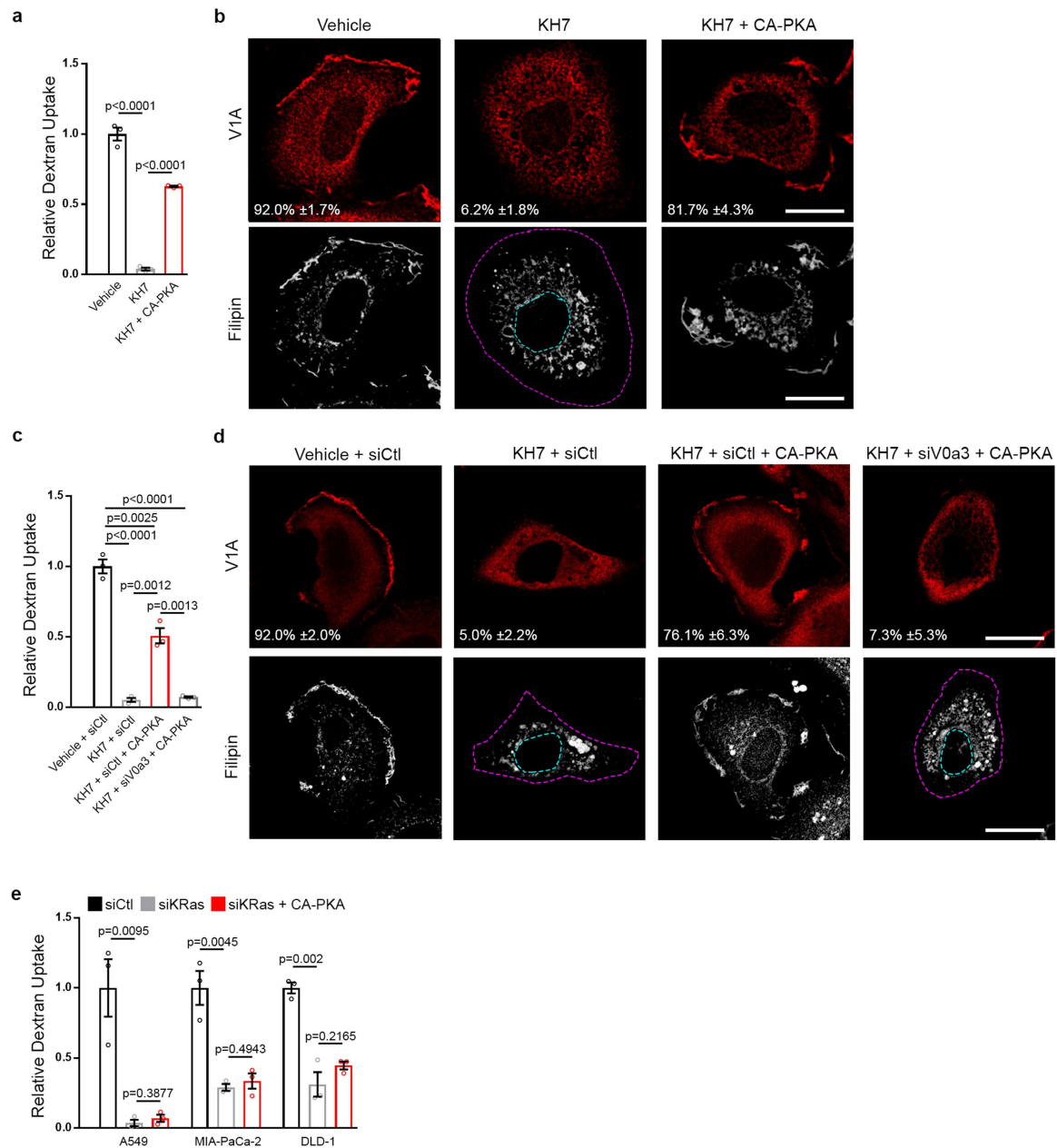
RAS-induced macropinocytosis. a–g, Effect of sAC (KH7) or PKA (H89)

inhibition on membrane cholesterol, RAC1 activation and RAC1 localization.

a, Fluorescence micrographs of HeLa HRAS(G12V) cells in indicated treatments with filipin labelling (top), membrane labelling with R-pre (middle), and merge of filipin and R-pre fluorescence micrographs with boxed areas enlarged beneath the image to show plasma membrane localization (bottom).

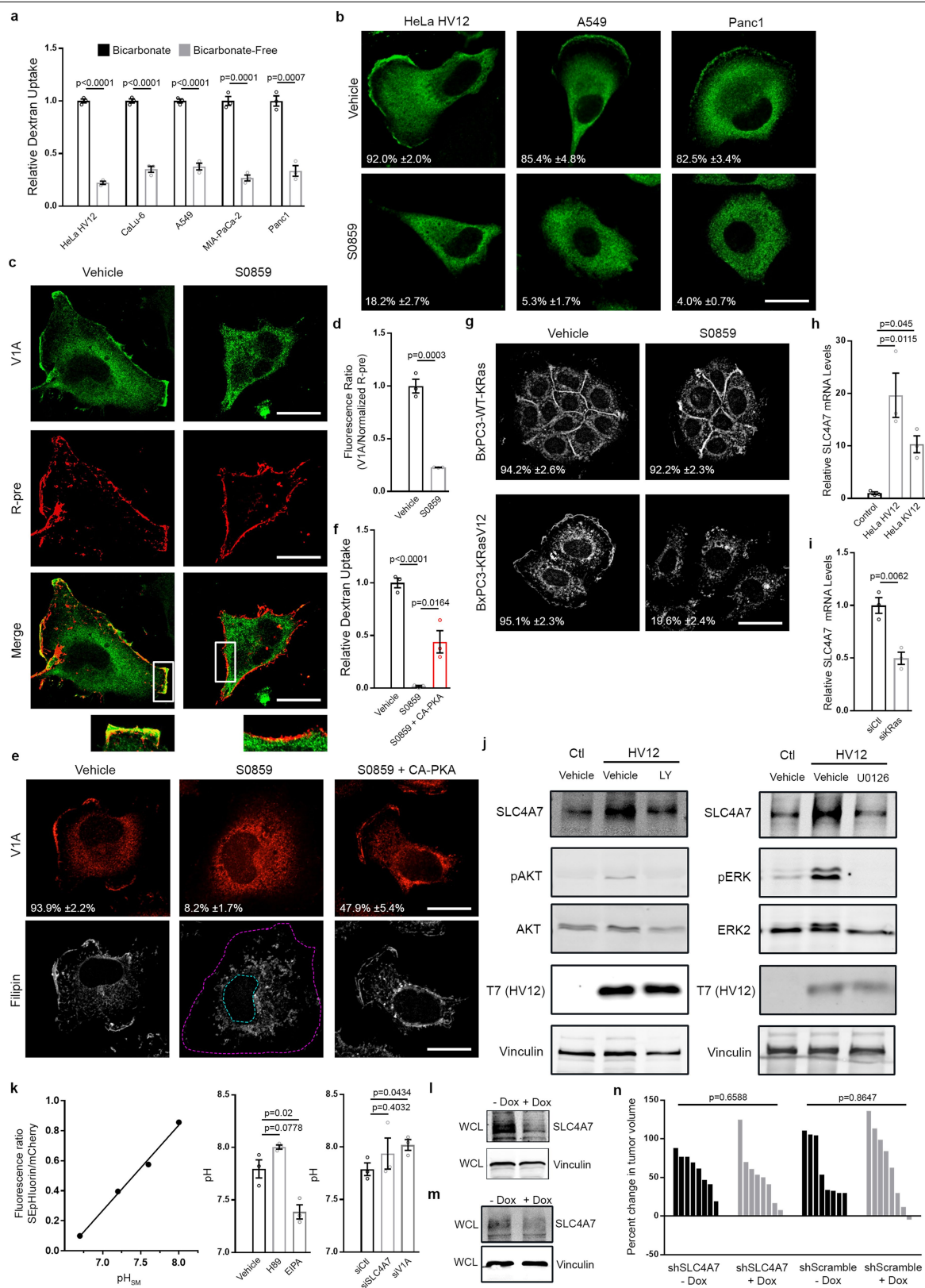
b, Quantification of the ratio of filipin to R-pre membrane localization with the indicated treatments in HeLa HRAS(G12V) cells. **c**, Immunoblot of endogenous RAC1 activity (GST-PBD, pull-down of RAC1-GTP; tubulin loading control) in HeLa control and HRAS(G12V) cells treated as indicated. **d**, Immunoblot of endogenous RAC1 in the plasma membrane fraction and whole-cell lysate in

HeLa control and HRAS(G12V) cells treated as indicated. **e**, Fluorescence micrographs of cholesterol (filipin) distribution in HeLa control cells with indicated treatments. **f**, Quantification of plasma membrane cholesterol in HeLa control cells with the indicated treatments. **g**, Fluorescence micrographs of cholesterol (filipin) distribution in BxPC-3 cells in the absence or presence of ectopically expressed KRAS(G12V) treated as indicated. **g**, Data are mean \pm s.e.m. representing the fraction of cells that display cholesterol plasma membrane localization. Images (**a**, **e**, **g**) and immunoblots (**c**, **d**) are representative of three biological replicates. Scale bars, 10 μ m. At least 20 (**b**), 50 (**f**) and 500 (**g**) cells were quantified in each biological replicate ($n = 3$); data are mean \pm s.e.m.; unpaired two-tailed Student's *t*-test. Gel source data for **c**, **d** are shown in Supplementary Fig. 1.



Extended Data Fig. 6 | PKA activation rescues RAS-induced macropinocytosis from sAC inhibition. **a, b**, Effect of sAC inhibition (KH7) and rescue (KH7 + CA-PKA) on macropinocytosis, V-ATPase localization and cholesterol distribution in HeLa HRAS(G12V) cells. **a**, Quantification of TMR-dextran uptake following indicated treatments. **b**, Fluorescence micrographs of V1A immunostaining and filipin labelling following indicated treatments. **c, d**, Effect of sAC inhibition (KH7), plasma membrane V-ATPase inhibition (siV0a3), and rescue (KH7 + CA-PKA + siV0a3) on macropinocytosis, V-ATPase localization and cholesterol distribution in HeLa HRAS(G12V) cells. **c**, Quantification of TMR-dextran uptake following indicated treatments.

d, Fluorescence micrographs of V1A immunostaining and filipin labelling following indicated treatments. **e**, Effect of KRAS inhibition (siKRAS) and rescue (siKRAS + CA-PKA) on macropinocytosis by quantification of TMR-dextran uptake in mutant RAS cell lines. Images in **b, d** are representative of three biological replicates. In **b, d**, dashed lines delineate the cell and nucleus; data are mean \pm s.e.m. representing the fraction of cells that display V1A plasma membrane localization. Scale bars, 10 μ m. At least 500 cells (**a, c, e**) were quantified in each biological replicate ($n = 3$); data are mean \pm s.e.m.; unpaired two-tailed Student's *t*-test.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | The SLC4 family is required for PKA-dependent, RAS-induced macropinocytosis. **a**, Quantification of TMR-dextran uptake in the absence (bicarbonate-free) or presence (bicarbonate) of extracellular bicarbonate in mutant RAS cells. **b**, Fluorescence micrographs of V1A immunostaining following treatment of mutant RAS cells with vehicle or pan-SLC4 inhibitor (S0859). **c**, Fluorescence micrographs of HeLa HRAS(G12V) cells in indicated treatments with V1A immunostaining (top), membrane labelling with R-pre (middle), and merge of V1A and R-pre fluorescence micrographs with boxed areas enlarged beneath the image to show plasma membrane localization (bottom). **d**, Quantification of the ratio of V1A to R-pre membrane localization with the indicated treatments in HeLa HRAS(G12V) cells from **c**. **e, f**, Effect of SLC4 inhibition (S0859) and rescue (S0859 + CA-PKA) in HeLa HRAS(G12V) cells. **e**, Fluorescence micrographs of V1A immunostaining and filipin labelling following indicated treatments. Dashed lines delineate the cell and nucleus. **f**, Quantification of TMR-dextran uptake. **g**, Fluorescence micrographs of cholesterol (filipin) distribution in BxPC-3 cells in the absence or presence of ectopically expressed KRAS(G12V) treated with SLC4 family inhibitor (S0859). **h**, mRNA levels of SLC4A7 expression in HeLa control, HRAS(G12V) or KRAS(G12V) cells. **i**, mRNA levels of SLC4A7 expression following KRAS knockdown in MIA-PaCa-2 cells. **j**, Effect of PI3K (LY294002, left) or MEK (U0126, right) inhibition on SLC4A7 expression in HeLa control

and HRAS(G12V) cells. Immunoblots of SLC4A7 expression from whole-cell lysate (vinculin loading control). p-AKT (left) and p-ERK (right) immunoblots show inhibition of pathways by the indicated treatments. **k**, Effect of PKA (H89), NHE (EIPA), SLC4A7 (siSLC4A7) and V1A (siV1A) inhibition on submembranous pH (pH_{sm}) in HeLa HRAS(G12V) cells transfected with SEpHluorin-mCherry construct (genetically encoded ratiometric pH probe that is targeted to the inner leaflet of the plasma membrane). Calibration curve of SEpHluorin-mCherry (line graph) was performed with K^+ nigericin buffer. Quantification of submembranous pH with H89 and EIPA treatment (bar graph, middle) or with knockdown of SLC4A7 and V1A (bar graph, right). **l**, Immunoblot of SLC4A7 (vinculin loading control) in MIA-PaCa2 cells with doxycycline-inducible SLC4A7 depletion. **m, n**, Effect of doxycycline-inducible SLC4A7 depletion in BxPC-3 cells on tumour growth. **m**, Immunoblot of SLC4A7 expression from whole-cell lysate (vinculin loading control). **n**, Waterfall plots of xenografts treated as shown relative to baseline. Each bar represents a tumour. Images (**b, c, e, g**), immunoblots (**j, l, m**), and mRNA levels (**h, i**) are representative of three biological replicates. **b, e, g**, Data are mean \pm s.e.m. representing the fraction of cells that display V1A (**b, e**) or cholesterol (**g**) plasma membrane localization. Scale bars, 10 μ m. At least 20 (**d**) and 500 (**a, b, e-g**) cells were quantified in each biological replicate ($n=3$); data are mean \pm s.e.m.; unpaired two-tailed Student's *t*-test. Gel source data for **j, l, m** are shown in Supplementary Fig. 1.

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see [Reporting Life Sciences Research](#). For further information on Nature Research policies, including our [data availability policy](#), see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study. For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

► Experimental design

1. Sample size

Describe how sample size was determined.

No statistical methods were used to predetermine sample sizes for in vitro experiments. Sample sizes were chosen in order to be able to perform statistical analyses, as is standard in the field. For in vivo experiments, the sample size was deemed sufficiently powered to detect a statistically significant and biologically relevant effect based on previous experiments using human pancreatic cell lines and xenograft model systems, which were utilized in this study.

2. Data exclusions

Describe any data exclusions.

No data were excluded from the analyses.

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

To verify reproducibility, experiments were performed using three biological replicates, unless clearly stated otherwise in the figure legend. All attempts at replication were successful.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

No randomization was performed for in vitro experiments. All in vitro experiments were carried out with appropriate internal negative and/or positive controls as indicated. For in vivo experiments, mice were randomized into groups of equal average tumour volumes.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

No blinding was performed for in vitro experiments. All in vitro experiments were carried out with appropriate internal negative and/or positive controls as indicated. Most results were validated by alternative techniques as described in this manuscript. For in vivo experiments, investigators were blinded once the mice were separated into experimental and control arms by the mice being given a coded number. During the experiment, one investigator measured the tumour volume and read the coded number to the second investigator who recorded the data for analysis.

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a Confirmed

- ☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
- ☐ ☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ A statement indicating how many times each experiment was replicated
- ☐ ☒ The statistical test(s) used and whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
- ☐ ☒ Test values indicating whether an effect is present
*Provide confidence intervals or give results of significance tests (e.g. *P* values) as exact values whenever appropriate and with effect sizes noted.*
- ☐ ☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
- ☐ ☒ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on [statistics for biologists](#) for further resources and guidance.

► Software

Policy information about [availability of computer code](#)

7. Software

Describe the software used to analyze the data in this study.

String 9.0 was used to identify functional clusters from screen hits. GraphPad Prism 6 was used for statistical analysis. Human mRNA raw data was analyzed using a GC-content background correction Robust Multi-array Average (RMA) algorithm (GC-RMA), performed in R: A language and environment for statistical computing. Macropinocytic Index was determined using ImageJ (ver. 1.50i). The ImageJ plugin, JACoP v2.033, was used to calculate the Mander's overlap colocalization coefficient of the v-ATPase with Rab7 for each cell analyzed.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* [guidance for providing algorithms and software for publication](#) provides further information on this topic.

► Materials and reagents

Policy information about [availability of materials](#)

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

All unique materials are readily available from the authors, except for where restricted by licensing agreements (for example, pTripz and any modifications thereof are explicitly prohibited from distribution by GE Healthcare). Data from the siRNA screen will be made available at NCBI PubChem.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

ATP6V1A: Abnova, H00000523-M02, clone 4F5. Validation, human: see ED1c (immunoblot); ED3b (IF).
 ATP6V0a3: Novus, nbp1-89333. Validation, human: in-house; siATP6V0a3 as shown in manuscript. Knockdown was tested by IB nbp1-89333.
 SLC4A7: Santa Cruz, sc-99633, L-15, Lot K1815. Validation immunoblotting, human: see4d.
 Vinculin: Sigma, V9264, clone hVIN-1. Validation: Used by hundreds of references as a loading control.
 FLAG: Sigma, F3165, clone M2. Validation: see 4d, detectable only upon expression of Flag-KV12.
 GFP: Cell Signaling, 2956, clone D5.1. Validated by ectopic expression of GFP-Rac1 in western blot.
 T7: Novagen (Millipore), 69522. Validation: see ED7j, detectable only upon expression of T7-HV12.
 Tubulin: Sigma, T5168, clone B-5-1-2. Validation: Used by hundreds of references as a loading control.
 Rac1: BD Transduction Laboratories, BD610650, clone 102/Rac1. Validation: in-house, published in Nimnual AS, etal 2008 Small GTPases.
 AKT: Cell Signaling, 2920, Clone 40D4. Validation: Cell Signaling by blocking peptide and recombinant protein.
 pAKT(S473): Cell Signaling, 4051 clone 587F11. Validation: immunoblotting, human: see ED7j, inhibition by LY294002.
 ERK2: EMD Millipore, 05-157, clone 1B3B9. Validation: immunoblotting, human: see ED7j, inhibition by U0126 (Dual IR channels with p-ERK on LiCor).
 pERK1/2: Cell Signaling, 4370, clone D13.14.4E. Validation: immunoblotting, human: see ED7j, inhibition by U0126.
 KRas: Santa Cruz, sc-30, clone F234. Validation: immunoblotting, human: see 4f.
 Rab7: Cell Signaling, 9367S, clone D95F2. Validation: Sapmaz, A, etal. 2019 Nature Comm 10: use siRab7, detect loss by IB using Cell Signaling 9367.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

Human cancer cell lines HeLa, CaLu-6, A549, MIA-PaCa-2, Panc-1, DLD-1, HCT-116, and BxPC-3 were obtained from the American Type Culture Collection (Manassas, VA).

b. Describe the method of cell line authentication used.

Cell lines are routinely authenticated in-house by cell morphology.

c. Report whether the cell lines were tested for mycoplasma contamination.

All cell lines used tested negative routinely for mycoplasma contamination by DAPI staining.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by [ICLAC](#), provide a scientific rationale for their use.

No commonly misidentified cell lines were used.

► Animals and human research participants

Policy information about [studies involving animals](#); when reporting animal research, follow the [ARRIVE guidelines](#)

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.

7 week old female homozygous NCr nude mice were used for animal studies.

Policy information about [studies involving human research participants](#)

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Study did not involve human research participants.

KRAS4A directly regulates hexokinase 1

<https://doi.org/10.1038/s41586-019-1832-9>

Received: 10 August 2018

Accepted: 23 October 2019

Published online: 11 December 2019

Caroline R. Amendola^{1,4}, James P. Mahaffey^{1,4}, Seth J. Parker¹, Ian M. Ahearn¹, Wei-Ching Chen², Mo Zhou¹, Helen Court¹, Jie Shi¹, Sebastian L. Mendoza¹, Michael J. Morten¹, Eli Rothenberg¹, Eyal Gottlieb³, Youssef Z. Wadghiri¹, Richard Possemato¹, Stevan R. Hubbard¹, Allan Balmain², Alec C. Kimmelman¹ & Mark R. Philips^{1*}

The most frequently mutated oncogene in cancer is *KRAS*, which uses alternative fourth exons to generate two gene products (*KRAS4A* and *KRAS4B*) that differ only in their C-terminal membrane-targeting region¹. Because oncogenic mutations occur in exons 2 or 3, two constitutively active *KRAS* proteins—each capable of transforming cells—are encoded when *KRAS* is activated by mutation². No functional distinctions among the splice variants have so far been established. Oncogenic *KRAS* alters the metabolism of tumour cells³ in several ways, including increased glucose uptake and glycolysis even in the presence of abundant oxygen⁴ (the Warburg effect). Whereas these metabolic effects of oncogenic *KRAS* have been explained by transcriptional upregulation of glucose transporters and glycolytic enzymes^{3–5}, it is not known whether there is direct regulation of metabolic enzymes. Here we report a direct, GTP-dependent interaction between *KRAS4A* and hexokinase 1 (HK1) that alters the activity of the kinase, and thereby establish that HK1 is an effector of *KRAS4A*. This interaction is unique to *KRAS4A* because the palmitoylation–depalmitoylation cycle of this *RAS* isoform enables colocalization with HK1 on the outer mitochondrial membrane. The expression of *KRAS4A* in cancer may drive unique metabolic vulnerabilities that can be exploited therapeutically.

We used affinity purification and mass spectrometry to analyse the proteins that interact with NRAS (Supplementary Table 1). This analysis identified HK1 (the ubiquitously expressed isozyme that initiates glucose metabolism); all three isoforms of the mitochondrial voltage-dependent anion channel (VDAC), which form complexes with HK1 on the outer mitochondrial membrane (OMM)⁶; and the ADP/ATP translocase 1 of the inner mitochondrial membrane, which associates with VDACs⁷. Co-immunoprecipitation of Flag-tagged NRAS and endogenous HK1 validated the interaction and revealed that it depends on GTP (Fig. 1a). Notably, the ability of HK1 to co-immunoprecipitate with *RAS* proteins depended on the *RAS* isoform: *KRAS4A* > NRAS >> HRAS > *KRAS4B*. Hexokinase 2 (HK2) has 73% sequence identity with HK1, shares all structural features, and is expressed in many cancer cells⁸. HK2 associated only with *KRAS4A* (Extended Data Fig. 1). Reciprocal co-immunoprecipitations that pulled down GFP-tagged HK1 or HK2 revealed exquisite isoform specificity, with affinity capture of only *KRAS4A* (Fig. 1b). The human colorectal cancer cell lines HT55 and GP5d both express relatively high levels of *KRAS4A*, but only GP5d cells contain an oncogenic *KRAS* mutation. Endogenous HK1 co-immunoprecipitated with endogenous *RAS* in lysates of GP5d but not HT55 cells (Fig. 1c).

Differential trafficking of *RAS* proteins is driven by post-translational modifications that include palmitoylation, which among the *KRAS* splice variants is unique to *KRAS4A*. Palmitoylation of *KRAS4A* on cysteine 180 in the C-terminal membrane-targeting region is required for efficient association with the plasma membrane¹. Because palmitoylation is reversible and short-lived⁹, palmitoylated *RAS*

proteins continuously cycle between membrane compartments¹⁰. Whereas a C186S mutation of Flag-tagged *KRAS4A*, which blocks prenylation and therefore all membrane association, completely blocked associations with haemagglutinin (HA)-tagged HK1 and HK2, a C180S mutation that blocks palmitoylation enhanced the associations (Fig. 2a). A similar result was obtained for endogenous HK1 and HK2 (Extended Data Fig. 2). Inhibition of palmitoylation with 2-bromopalmitate (2-BP)¹¹ enhanced the association of endogenous *RAS* with HK1 (Fig. 2b). Thus, whereas prenylation was required for the interaction of *KRAS4A* with hexokinases, palmitoylation—which drives efficient association with the plasma membrane—negatively regulated the interaction.

KRAS4A and HK1 interact on mitochondria

Because HK1 and HK2 are targeted to the OMM¹², our data suggest that depalmitoylated *KRAS4A* may have affinity for the OMM, which would thereby support the association between *KRAS4A* and hexokinases. To test this hypothesis, we co-expressed mCherry-tagged *KRAS4A*(G12V), with or without mutation of cysteine 180 to serine (C180S) to eliminate palmitoylation, with GFP that was targeted to the OMM by extension with the mitochondrial targeting sequence of HK1 (amino acids 1–16). Whereas mCherry–*KRAS4A*(G12V) was predominantly observed on the plasma membrane, and to a lesser degree on intracellular vesicles, mCherry–*KRAS4A*(G12V/C180S) colocalized with mitochondrially targeted GFP on the OMM (Extended Data Fig. 3a, b). By contrast, although depalmitoylated NRAS also accumulated on endomembranes¹, these

¹Perlmutter Cancer Center, NYU School of Medicine, New York, NY, USA. ²Helen Diller Family Comprehensive Cancer Center, University of California at San Francisco School of Medicine, San Francisco, CA, USA. ³Technion Israel Institute of Technology, Haifa, Israel. ⁴These authors contributed equally: Caroline R. Amendola, James P. Mahaffey. *e-mail: mark.philips@nyulangone.org

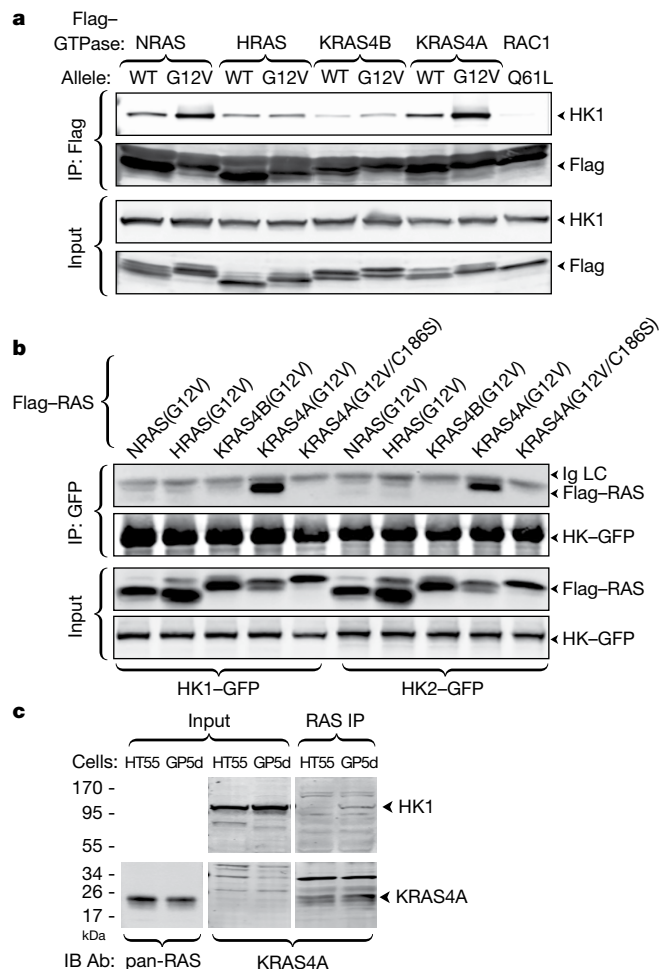


Fig. 1 | KRAS4A binds to HK1 and HK2 in a GTP- and prenylation-dependent manner. **a**, The indicated Flag-tagged RAS constructs (with or without G12V mutations) were expressed in HeLa cells and immunoprecipitated (IP) with anti-Flag beads. Blots were probed for endogenous HK1. Flag–RAC1(Q61L) served as the negative control. WT, wild type. Data are representative of four independent experiments ($n = 4$). **b**, GFP-tagged HK1 or HK2 were co-expressed with the indicated Flag-tagged RAS constructs, immunoprecipitated with anti-GFP beads, and the blots were probed with anti-Flag antibody. KRAS4A(G12V/C186S) is not prenylated and therefore cannot associate with membranes ($n = 3$). Ig LC, immunoglobulin light chain. **c**, Co-immunoprecipitation of endogenous HK1 with endogenous KRAS4A (captured by the monoclonal antibody Y13-259) in colorectal tumour cells with oncogenic KRAS (GP5d) but not those with wild-type KRAS (HT55). IB, immunoblotting.

did not include the OMM (Extended Data Fig. 3c). Super-resolution immunofluorescence microscopy confirmed colocalization of palmitoylation-deficient KRAS4A and HK1 on the OMM (Fig. 2c). KRAS4B did not colocalize with HK1 on the OMM (Extended Data Fig. 3d). HCT-15 cells are KRAS-mutant colorectal cancer cells that express KRAS4A at relatively high levels. We detected endogenous KRAS4A on mitochondria that were rapidly isolated from these cells (Extended Data Fig. 3e), and the mitochondria-associated pool increased by 63% when cells were pre-treated with 2-BP. Thus, endogenous depalmitoylated KRAS4A colocalizes with HK1 on the OMM.

To determine whether the interaction between KRAS4A and HK1 is direct, we used a fully recombinant system. Bacterially expressed KRAS4A was affinity captured by the RAS-binding domain (RBD) of RAF1 or HK1, each of which was fused to glutathione *S*-transferase (GST), but not by GST alone (Fig. 3a). Notably, the interaction was observed only when KRAS4A was loaded with GTP. The GTP dependence suggests that

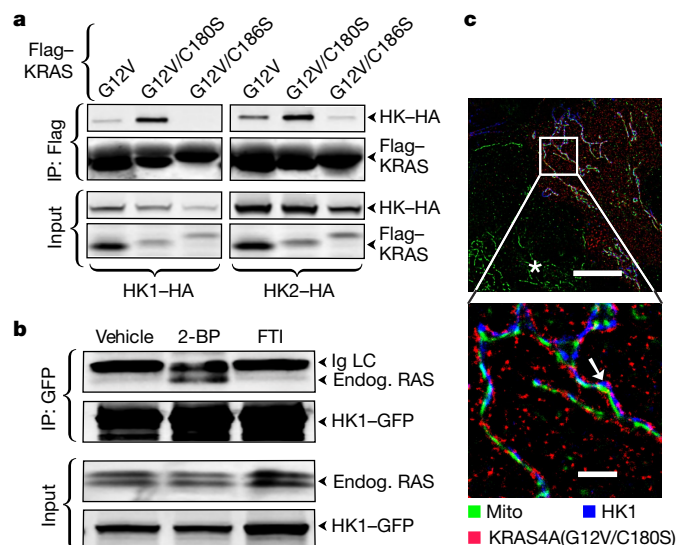


Fig. 2 | Depalmitoylated KRAS4A interacts with HK1 on the OMM. **a**, HA-tagged HK1 or HK2 were co-expressed in HEK293 cells with the indicated Flag-tagged KRAS4A constructs. KRAS was immunoprecipitated with anti-Flag beads and analysed by immunoblot probed with anti-HA and anti-Flag antibodies ($n = 3$). **b**, GFP-tagged HK1 was expressed in HCT-116 cells treated with vehicle, 25 μ M 2-BP or 20 μ M farnesyl transferase inhibitor (FTI). HK1-GFP was immunoprecipitated and the blots were probed with an anti-pan-RAS antibody ($n = 2$). **c**, U2OS cells expressing Flag-tagged KRAS4A(G12V/C180S) and HK1-GFP were treated with MitoTracker, fixed, stained for Flag and GFP and imaged by super-resolution stochastic optical reconstruction microscopy (STORM). Asterisk indicates an untransfected cell; arrow indicates colocalization of KRAS4A and HK1 on the OMM. Mito, mitochondria. The image is representative of $n = 5$ independent experiments. Scale bars, 5 μ m (top); 1 μ m (bottom).

the G domain of KRAS4A interacts with HK1. Given that the G domains of all four RAS proteins are nearly identical, we hypothesized that without membrane targeting there should be no isoform specificity. Indeed, in an *in vitro* assay driven by mass action, recombinant, non-prenylated KRAS4B—but not RAC2—behaved like KRAS4A and interacted with GST-tagged HK1 in a GTP-dependent manner (Fig. 3b). The GTP-dependent, direct protein–protein interaction suggests that HK1 contains a structural analogue of an RBD. Examination of published HK1 structures revealed a surface-exposed helix–loop–sheet motif that is characteristic of RBDs¹³ (Extended Data Fig. 4). We expressed the isolated putative RBD region (HK1 amino acids 76–206) extended with the HK1 mitochondrial targeting sequence and tagged with HA and found that it was as effective at pulling down KRAS4A(G12V) as the N-terminal kinase domain of HK1 (Fig. 3c). These data suggest that the region of HK1 that comprises amino acids 76–206 functions as an RBD.

The binding affinities of RBDs for RAS vary widely¹³. The two-dimensional surface of the mitochondrion might promote the interaction of KRAS4A with HK1 even if the affinity is relatively weak. To confirm this model, we artificially targeted KRAS4B and HRAS (constitutively bound to GTP) to the OMM by extending them at their N termini with the mitochondrial targeting sequence of HK1. Whereas HK1 co-immunoprecipitated with neither Flag–KRAS4B(G12V) nor Flag–HRAS(G12V), a robust association was detected between HK1 and both mitochondrially targeted Flag–KRAS4B(G12V) and Flag–HRAS(G12V) (Fig. 3d). By contrast, GTP-loaded RAC1(Q61L) that was targeted to the OMM in the same way did not associate with HK1. Conversely, removing the mitochondrial targeting sequence from HK1 diminished its ability to interact with KRAS4A (Extended Data Fig. 3d). Thus, the isoform

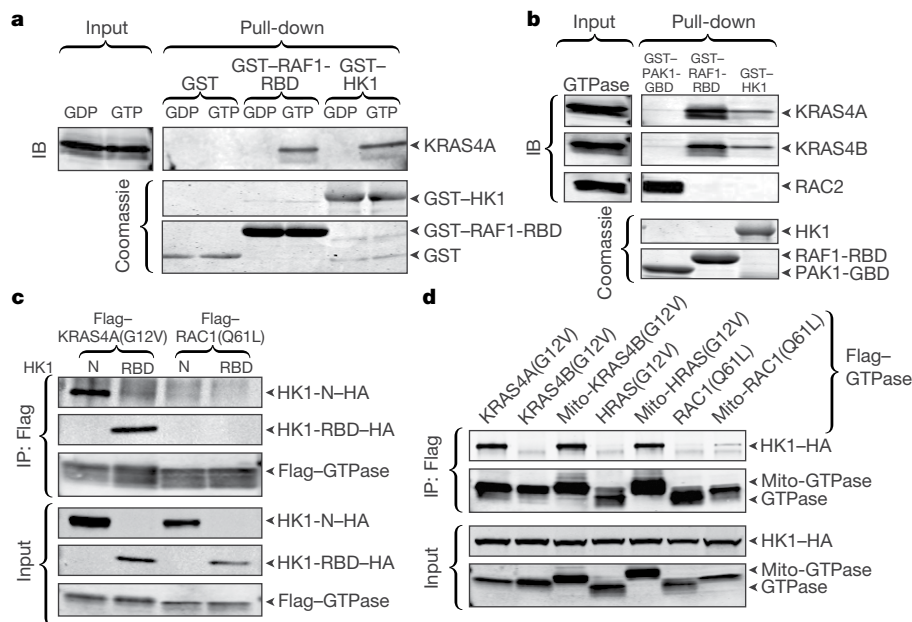


Fig. 3 | The interaction between KRAS4A and HK1 is direct, dependent on GTP, mediated by an RBD-like region of HK1 and requires OMM localization.

a, Recombinant KRAS4A was loaded with GDP or GTPγS and then incubated with GST alone, the GST-tagged RBD of RAF1 (GST-RAF1-RBD) or GST-tagged HK1 (GST-HK1). Affinity capture was assessed by immunoblot for RAS. GST loading is shown by a Coomassie-stained gel ($n = 3$). **b**, Recombinant KRAS4A, KRAS4B and RAC2 were loaded with GTPγS and then incubated with GST fused to the GTPase binding domain (GBD) of PAK1 (GST-PAK1-GBD), GST-RAF1-RBD or GST-HK1. Affinity capture was assessed by immunoblot for RAS or RAC2

($n = 2$). **c**, The HA-tagged N-terminal domain of HK1 or the isolated putative RBD (amino acids 76–206) extended with the mitochondrial targeting sequence of HK1, were expressed with Flag-tagged KRAS4A(G12V) or RAC1(Q61L) and Flag immunoprecipitates were blotted as indicated ($n = 3$). **d**, HA-tagged HK1 was co-expressed in HEK293 cells with Flag-tagged KRAS4A(G12V) or KRAS4B(G12V), HRAS(G12V) or RAC1(Q61L), with or without N-terminal extension with the mitochondrial targeting sequence of HK1 (Mito). The Flag-tagged GTPases were immunoprecipitated and binding to HK1 was assessed with anti-HA antibody. $n = 4$.

specificity of the interaction with regard to KRAS4A is driven by its membrane-targeting sequence, which permits colocalization with HK1 on the OMM.

KRAS4A blocks allosteric inhibition of HK1

To determine whether KRAS4A directly affects the enzymatic activity of HK1, we studied recombinant, GST-tagged enzymes associated with glutathione–agarose beads (to mimic the OMM and allow multimerization)¹⁴. Neither recombinant, GTP-loaded KRAS4A nor RAC2 affected the kinetics of full-length HK1 (Extended Data Fig. 5a). Both HK1 and HK2 consist of two tandem kinase domains. Whereas both domains are catalytic in HK2¹⁵, only the C-terminal domain of HK1 can phosphorylate glucose. The HK1 N-terminal domain provides allosteric feedback inhibition by the product of the reaction, glucose-6-phosphate¹⁶. As expected, 2-deoxyglucose (2-DG) slowed the reaction (Fig. 4a–c). 2-DG is both a competitive and a non-competitive inhibitor of HK1: aside from directly competing with glucose it is also converted by hexokinases to 2-deoxyglucose-6-phosphate, which cannot be further metabolized and therefore accumulates and acts as an analogue of glucose-6-phosphate to mediate allosteric feedback inhibition¹⁶. Whereas recombinant RAC2 had no effect on the 2-DG-inhibited reaction, recombinant KRAS4A partially reversed 2-DG-mediated inhibition (Fig. 4a). By contrast, KRAS4A did not reverse 2-DG-mediated inhibition of either the isolated C-terminal kinase domain of HK1 (Fig. 4b) or full-length HK2 (Fig. 4c), which suggests that the action of KRAS4A on full-length HK1 is through the allosteric site. Consistent with this interpretation, KRAS4A restored the velocity of enzyme-catalysed reaction at infinite concentration of substrate (V_{\max}) that was diminished by 2-DG, but had relatively little effect on the Michaelis constant (K_m) (Extended Data Fig. 5b). The mitigating effect of KRAS4A on the allosteric inhibition induced by 2-DG

demonstrates that the direct protein–protein interaction between KRAS4A and HK1 has a functional consequence.

KRAS4A enhances glycolytic flux

To correlate these in vitro results with cellular glucose metabolism, we expressed oncogenic forms of the *KRAS* splice variants in HEK293 cells. As expected, both oncogenic KRAS4A and KRAS4B enhanced the glucose consumption of these cells; however, when expressed at similar levels, KRAS4A had twice the effect of KRAS4B (Fig. 4d). When HK1, but not HK2, was silenced, this differential was lost, which suggests that the mechanism by which KRAS4A exceeds KRAS4B in promoting glucose consumption requires HK1. This result was confirmed by metabolic analysis of Flp-In T-REx 293 cells that were induced with doxycycline to express equivalent levels of KRAS4A(G12V), KRAS4A(G12V/C180S) or KRAS4B(G12V). Both glucose consumption (Extended Data Fig. 6a) and basal extracellular acidification rate (ECAR) (Extended Data Fig. 6b) were increased in cells that expressed KRAS4A(G12V) relative to those that expressed KRAS4B(G12V). Notably, the palmitoylation-deficient mutant KRAS4A(G12V/C180S) was more potent than KRAS4A(G12V), which indicates that there is an association between localization to the OMM and enhanced glucose consumption. Whereas KRAS4A and KRAS4B were equivalent in inducing phosphorylation of MEK, ERK and AKT, KRAS4A(G12V/C180S) was less effective (Extended Data Fig. 6c, d). Moreover, none of the *KRAS* proteins increased the levels of HK1 and HK2. Thus, increased glucose metabolism was dissociated in this system from *KRAS* signalling down the MAPK and PI3K pathways and from transcriptional activation of hexokinases.

We performed the converse experiments by targeting the 4A exon of *KRAS* with CRISPR–Cas9 in two human tumour cell lines that contain oncogenic *KRAS* mutations: A549 (lung) and SUIT2 (pancreas). In

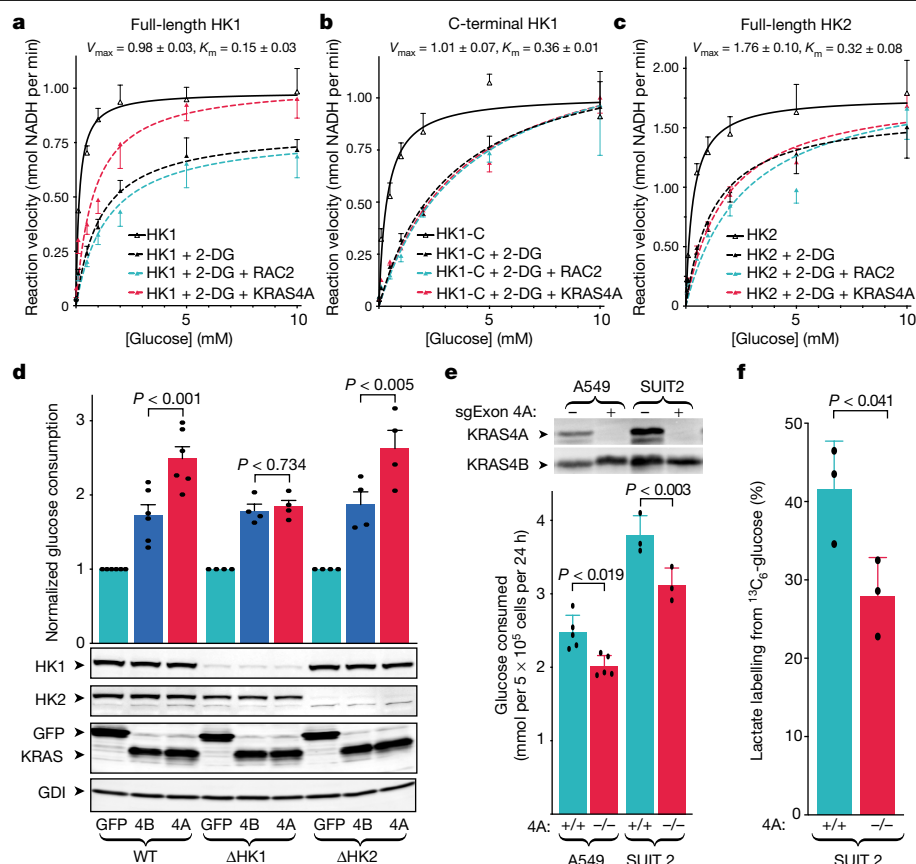


Fig. 4 | KRAS4A increases HK1 activity in vitro and in vivo. **a–c**, Activity of recombinant full-length HK1 (**a**), the C-terminal kinase domain of HK1 (**b**) or full-length HK2 (**c**). Reaction velocities are plotted (mean \pm s.e.m.) as a function of glucose concentration. Velocities with and without 2-DG (20 mM) are shown, with or without the addition of recombinant, GTP-loaded RAC2 or KRAS4A. Plots combine independent assays ($n = 4$ (**a**); $n = 3$ (**b**, **c**)). **d**, Wild-type ($n = 6$), HK1-deficient (Δ HK1; $n = 4$) or HK2-deficient (Δ HK2; $n = 4$) HEK293 cells were transfected with Flag-tagged GFP, KRAS4B(G12V) or KRAS4A(G12V), and the glucose consumption over 24 h (mean \pm s.e.m.) was determined. The

immunoblot shows relative expression. **e**, Glucose consumption (mean \pm s.e.m.) of parental A549 ($n = 5$) and SUIT2 ($n = 3$) human cancer cells and those in which exon 4A of *KRAS* was disrupted with CRISPR–Cas9 (sgExon 4A). The immunoblot in **e** shows the absence of KRAS4A. **f**, Incorporation of the ^{13}C label from glucose into lactate (mean \pm s.e.m.) over 15 min in SUIT2 cells with or without exon 4A of *KRAS* ($n = 3$). Significance in **d–f** was determined by two-sided Student's *t*-test (paired in **d**, **e**; unpaired in **f**). Growth rates of the two genotypes were equal over the 24-h assessment.

both cases, disrupting exon 4A reduced glucose consumption (Fig. 4e, Extended Data Fig. 7a), lactate secretion (Extended Data Fig. 7b) and ECAR (Extended Data Fig. 8). Glucose consumption was restored by forced expression of KRAS4A (Extended Data Fig. 7a). To determine whether these results reflect alterations in glycolytic flux, we labelled glucose with the ^{13}C stable isotope, and found that disruption of exon 4A diminished the conversion of glucose into lactate (Fig. 4f, Extended Data Fig. 7c). In addition, cells in which the 4A exon was disrupted (*KRAS4A*^{−/−}) were more sensitive to growth inhibition by 2-DG than were *KRAS4A*^{+/+} cells in both A549 and SUIT2 cell lines (Extended Data Fig. 7d, e)—consistent with diminished glycolytic capacity. We measured the uptake of glucose by cells in vivo using ^{18}F -deoxyglucose positron emission tomography (^{18}F -FDG-PET) in xenograft tumours that were generated with SUIT2 (Extended Data Fig. 7f) or A549 (Extended Data Fig. 7g) cells with the genotypes *KRAS4A*^{+/+} or *KRAS4A*^{−/−}. Tumours that expressed KRAS4A (*KRAS4A*^{+/+}) took up glucose at a faster rate per unit volume than did *KRAS4A*^{−/−} tumours.

Because the regulation of HK1 by KRAS4A appears to be stoichiometric (as is the case for all RAS effectors¹⁷) we sought to determine the relative abundance of KRAS4A and HK1. Using quantitative immunoblotting, we determined that A549 and HCT-15 cells contain on average 7×10^5 and 2×10^5 molecules of KRAS, and 2×10^5 and 1.5×10^5 molecules of HK1, respectively (Extended Data Fig. 9). Given that the KRAS4A variant makes up 15–50% of KRAS expression¹ and HK1 functions as

a homodimer, we conclude that there is sufficient KRAS4A in these cells to have a physiologically meaningful effect on catalytic activity.

Discussion

An effector of a small GTPase such as RAS must meet three requirements: binding must (1) be direct; (2) occur only with the GTP-bound form; and (3) alter the activity of the effector. Our data demonstrate that HK1 meets these requirements and is therefore an effector of KRAS4A. HK1 is unique in two ways: first, it is an effector of a GTPase that is also a metabolic enzyme; and second, it discriminates between the two *KRAS* splice variants. Although—like all effector interactions—binding takes place through the G domain of KRAS4A, it is the distinct subcellular trafficking of KRAS4A and KRAS4B that establishes the differential engagement. We and others have previously reported that RAS signalling is compartmentalized by subcellular trafficking^{18–20}. KRAS4A signalling to HK1 on the OMM is another clear example of this phenomenon.

The metabolic rewiring of human tumours by oncogenic *KRAS* is probably a consequence of many factors, which include increased expression of transporters and enzymes^{8,21}. Our data suggest that direct regulation of HK1 by KRAS4A represents an additional mechanism. In support of this model (Extended Data Fig. 10), whereas maximal MAPK signalling does not correlate with KRAS expression, the ability of oncogenic *KRAS* to upregulate glycolysis depends on the copy number of mutant *KRAS*⁵.

Distinct roles in metabolism may be among the reasons for the persistence of alternative *KRAS* splicing throughout vertebrate evolution. Preferential expression of *KRAS4A* in the gastrointestinal tract²² may be due to the specific metabolic requirements of these tissues. HK1 and phosphofructokinase have been shown to be the glycolytic enzymes that control flux through the glycolytic pathway²³, which suggests that direct regulation of HK1 by *KRAS4A* would have a substantial effect. Moreover, because one of the desired outcomes of metabolic rewiring from the standpoint of sustaining rapid tumour growth is diverting glucose into the pentose phosphate pathway²⁴, HK1 is an ideal regulatory node. Metabolic reprogramming that is mediated by *KRAS* has long been seen as a potential vulnerability in cancer, and both HK1 and HK2 inhibitors have been investigated in this regard²⁵. Our findings suggest that efforts along these lines may prove most effective in tumours that express relatively high levels of *KRAS4A*, and that a better understanding of the regulation of *KRAS* splicing and *KRAS4A* palmitoylation may reveal new modes of therapy.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1832-9>.

1. Tsai, F. D. et al. K-Ras4A splice variant is widely expressed in cancer and uses a hybrid membrane-targeting motif. *Proc. Natl Acad. Sci. USA* **112**, 779–784 (2015).
2. Voice, J. K., Klemke, R. L., Le, A. & Jackson, J. H. Four human Ras homologs differ in their abilities to activate Raf-1, induce transformation, and stimulate cell motility. *J. Biol. Chem.* **274**, 17164–17170 (1999).
3. Kimmelman, A. C. Metabolic dependencies in RAS-driven cancers. *Clin. Cancer Res.* **21**, 1828–1834 (2015).
4. Ying, H. et al. Oncogenic Kras maintains pancreatic tumors through regulation of anabolic glucose metabolism. *Cell* **149**, 656–670 (2012).
5. Kerr, E. M., Gaude, E., Turrell, F. K., Frezza, C. & Martins, C. P. Mutant Kras copy number defines metabolic reprogramming and therapeutic susceptibilities. *Nature* **531**, 110–113 (2016).
6. Gottlob, K. et al. Inhibition of early apoptotic events by Akt/PKB is dependent on the first committed step of glycolysis and mitochondrial hexokinase. *Genes Dev.* **15**, 1406–1418 (2001).
7. Vyssokikh, M. Y. & Brdiczka, D. The function of complexes between the outer mitochondrial membrane pore (VDAC) and the adenine nucleotide translocase in regulation of energy metabolism and apoptosis. *Acta Biochim. Pol.* **50**, 389–404 (2003).
8. Patra, K. C. et al. Hexokinase 2 is required for tumor initiation and maintenance and its systemic deletion is therapeutic in mouse models of cancer. *Cancer Cell* **24**, 213–228 (2013).
9. Ahearn, I., Zhou, M. & Philips, M. R. Posttranslational modifications of RAS proteins. *Cold Spring Harb. Perspect. Med.* **8**, a031484 (2018).
10. Rocks, O. et al. The palmitoylation machinery is a spatially organizing system for peripheral membrane proteins. *Cell* **141**, 458–471 (2010).
11. Jennings, B. C. et al. 2-Bromopalmitate and 2-(2-hydroxy-5-nitro-benzylidene)-benzo[b]thiophen-3-one inhibit DHHC-mediated palmitoylation in vitro. *J. Lipid Res.* **50**, 233–242 (2009).
12. John, S., Weiss, J. N. & Ribalet, B. Subcellular localization of hexokinases I and II directs the metabolic fate of glucose. *PLoS One* **6**, e17674 (2011).
13. Wohlgemuth, S. et al. Recognizing and defining true Ras binding domains I: biochemical analysis. *J. Mol. Biol.* **348**, 741–758 (2005).
14. Zhu, A., Romero, R. & Petty, H. R. An enzymatic colorimetric assay for glucose-6-phosphate. *Anal. Biochem.* **419**, 266–270 (2011).
15. Tsai, H. J. & Wilson, J. E. Functional organization of mammalian hexokinases: both N- and C-terminal halves of the rat type II isozyme possess catalytic sites. *Arch. Biochem. Biophys.* **329**, 17–23 (1996).
16. Sebastian, S., Wilson, J. E., Mulichak, A. & Garavito, R. M. Allosteric regulation of type I hexokinase: a site-directed mutational study indicating location of the functional glucose 6-phosphate binding site in the N-terminal half of the enzyme. *Arch. Biochem. Biophys.* **362**, 203–210 (1999).
17. Rajalingam, K., Schreck, R., Rapp, U. R. & Albert, S. Ras oncogenes and their downstream targets. *Biochim. Biophys. Acta* **1773**, 1177–1195 (2007).
18. Chiu, V. K. et al. Ras signalling on the endoplasmic reticulum and the Golgi. *Nat. Cell Biol.* **4**, 343–350 (2002).
19. Mor, A. & Philips, M. R. Compartmentalized Ras/MAPK signaling. *Annu. Rev. Immunol.* **24**, 771–800 (2006).
20. Aran, V. & Prior, I. A. Compartmentalized Ras signaling differentially contributes to phenotypic outputs. *Cell. Signal.* **25**, 1748–1753 (2013).
21. Yun, J. et al. Glucose deprivation contributes to the development of *KRAS* pathway mutations in tumor cells. *Science* **325**, 1555–1559 (2009).
22. Newlaczyl, A. U., Coulson, J. M. & Prior, I. A. Quantification of spatiotemporal patterns of Ras isoform expression during development. *Sci. Rep.* **7**, 41297 (2017).
23. Tanner, L. B. et al. Four key steps control glycolytic flux in mammalian cells. *Cell Syst.* **7**, 49–62 (2018).
24. Vander Heiden, M. G., Cantley, L. C. & Thompson, C. B. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science* **324**, 1029–1033 (2009).
25. Hay, N. Reprogramming glucose metabolism in cancer: can it be exploited for cancer therapy? *Nat. Rev. Cancer* **16**, 635–649 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Cell lines, culture and transfection

The cell lines A549, COS1, HCT-15, HCT-116, HEK293, HeLa, SUIT2 and U2OS were from ATCC, which validated each line by short tandem repeat profiling. Flp-In T-REx293 cells were from Thermo Fisher Scientific and were validated by successful insertion of a gene of interest upon ectopic expression of Flp recombinase. *KRAS4A*^{-/-} A549 and SUIT2 cells were generated by introducing an indel into *KRAS* exon 4A with CRISPR-Cas9, and *HK1*- and *HK2*-deficient HEK293 cells were generated by targeting the second and third exons of *HK1* and *HK2*, respectively, with the same strategy. All cell lines were maintained in Dulbecco's modified Eagle medium (DMEM), except for HCT-15 cells, which were grown in RPMI medium. All media were supplemented with 10% FBS and 1× penicillin and streptomycin. Transgene expression was induced in Flp-In T-REx 293 cells with 0.75 mg ml⁻¹ treatment with doxycycline for 24 h. Serum starving was performed with medium containing 0.1–1.0% FBS. All cell lines were transfected with Lipofectamine 3000 reagent (Thermo Fisher Scientific) according to the manufacturer's protocol.

Recombinant proteins

All GST-tagged proteins (full-length HK1, C-terminal HK1, full-length HK2, PAK1-GBD, RAF1-RBD and free GST) and 6×His-tagged RAC2 were produced in the BL21 strain of *Escherichia coli* and purified in HEPES buffer (20 mM HEPES, pH 7.3, 150 mM NaCl, 5 mM MgCl₂ and 0.2% TritonX-100). HK1 proteins were induced with 0.75 mM IPTG at 18 °C for 24 h; all other constructs were induced at 37 °C for 3 h. *E. coli* were lysed in HEPES buffer containing 2% TritonX-100 by sonication, and GST-tagged proteins were isolated from clarified lysate using glutathione-agarose beads (Sigma-Aldrich). Recombinant RAS proteins were provided by W. Gillette. For GDP- and GTP-loading experiments, RAS proteins were incubated at 4 °C in HEPES buffer supplemented with 10 mM EDTA and a 10-molar excess (with respect to RAS) of GDP or GTPγS (Sigma-Aldrich) for 15 min to dissociate bound nucleotide. Then, 1 M MgCl₂ was added to the solution and it was incubated for 30 min at 4 °C to allow the rebinding of nucleotides.

Immunoprecipitation, in vitro binding and western blotting

Cells were lysed in co-immunoprecipitation buffer (50 mM Tris, pH 7.5, 150 mM NaCl, 10 mM NaF, 1 mM EDTA and 1% NP40) supplemented with 1× protease inhibitor tablet (Roche) and 1× PhosStop tablet (Sigma-Aldrich). Immunoprecipitation was performed in cleared cellular lysate using anti-Flag- (Sigma-Aldrich), anti-GFP- (MBL Life Science) or Y13-259 anti-RAS- (Sigma-Aldrich) conjugated agarose beads for 1–3 h at 4 °C. For in vitro binding assays, 0.5 µg of GST-tagged bait proteins bound to glutathione-agarose beads were incubated with 1.0 µg of the indicated recombinant GTPases for 1 h at 4 °C.

After washing beads with co-immunoprecipitation or HEPES buffer, bound proteins were eluted with SDS sample buffer and assayed by western blotting using 4–20% SDS-PAGE gels (Invitrogen) and PVDF membranes (Millipore). The following antibodies were used for western blotting: anti-Flag (Sigma-Aldrich, F7425, lot 085M4774V, 1:2,000), anti-GFP (Thermo Fisher Scientific, A-6455, lot 1826342, 1:2,000), anti-HK1 (CST, 2024, clone C35C4, lot 3, 1:2,000), anti-HK2 (CST, 2867, clone C64G5, lot 3, 1:2,000), anti-SDHA (CST, I1998, clone D6J9M, lot 2, 1:1,000), anti-pMEK (CST, 9121, lot 47, 1:1,000), anti-tMEK (CST, 4694, clone L38C12, 1:1,000), anti-phosphorylated (p)ERK (CST, 9106, clone E10, 1:1,000), anti-pAKT (CST, 9271, clone S473, lot 14, 1:1,000), anti-tAKT (CST, 9272, lot 27, 1:1,000), anti-pan-RAS (CalBiochem, OP40, clone Ab-3, lot D00119097, 1:2,000), anti-fibrillarin (Santa Cruz

Biotechnology, sc-374022, clone G-8, lot G0116, 1:1,000), anti-EEA1 (Santa Cruz Biotechnology, sc-137130, clone G-4, lot F2716, 1:1,000), anti-F₁-ATPase (Santa Cruz Biotechnology, sc-514419, clone C-12, lot E1016, 1:1,000), anti-total (t)ERK (Santa Cruz Biotechnology, sc-94, clone K-23, 1:1,000), anti-Rho GDI (Santa Cruz Biotechnology, sc-360, clone A-20, lot J0313, 1:5,000), anti-HA (Santa Cruz Biotechnology, sc-7392, clone F-7, lot K3012, 1:1,000), anti-RAC2 (Abcam, 130415, lot GR310183-1, 1:1,000) and anti-KRAS4A, a polyclonal rabbit antibody that was developed by our laboratory¹ and licensed for commercial distribution (Sigma-Aldrich, ABC1442, 1:500).

Mass spectrometry

Affinity-purified proteins were reduced, alkylated and loaded onto an SDS-PAGE gel to remove any detergents and reagents that were incompatible with liquid chromatography-mass spectrometry (LC-MS). The gel plugs were excised, de-stained and subjected to proteolytic digestion with trypsin. The resulting peptides were extracted, desalted and an aliquot was analysed with LC-MS coupled to a Thermo Fisher Scientific Orbitrap QExactive Mass Spectrometer operated in data-dependent mode as previously described²⁶. The data were searched against a UniProt human database, using Sequest within Proteome Discoverer.

Cellular and biochemical assays

Hexokinase activity assays were performed using the Hexokinase Assay Kit (BioVision), and assays that included 2-DG as an HK1 inhibitor were performed with 20 mM 2-DG (Sigma-Aldrich). In vitro glucose uptake assays were performed by culturing 5 × 10⁵ cells in 2 ml of low-serum, 5 mM glucose DMEM for 24 h. After culture, the medium was removed, cells and debris were cleared by centrifugation and the remaining glucose in the medium was measured using Infinity Glucose Oxidase Liquid Stable Reagent. Mitochondria were isolated from the post-nuclear supernatants of homogenates of cells disrupted by nitrogen cavitation as described previously²⁷ using a Mitochondria Isolation Kit, which relies on immunoaffinity purification with magnetic beads conjugated with anti-TOM70 antibodies (Miltenyi Biotec). The inhibition of cell growth by 2-DG was tested by growing cells in different concentrations of 2-DG and counting cells after 24 h of treatment with a Beckman Coulter cell counter.

Microscopy

Live-cell imaging was performed on cells transiently transfected in 35-mm plates that incorporate a well with a coverslip of 1.5 thickness (MatTek) with an inverted Zeiss 800 laser scanning confocal microscope. Zeiss Zen Blue 2 software (v.2.1) was used for data acquisition and analysis.

Super-resolution imaging by STORM

Cells were stained with MitoTracker, then fixed (3% PFA, 0.2% glutaraldehyde in PBS) and permeabilized (0.1% Triton X100 in PBS). Glycine (0.4 M) was added for 20 min to quench the glutaraldehyde. Cells were incubated in blocking buffer (2% glycine, 2% BSA, 0.2% gelatin and 50 mM NH₄Cl in PBS) overnight at 4 °C. Flag-KRAS4A and Flag-KRAS4A(G12V) were labelled with rabbit anti-Flag primary antibody (Thermo Fisher Scientific), then incubated with secondary antibody (goat anti-rabbit conjugated to Alexa Fluor 647; Abcam). HK1-GFP was stained with rabbit anti-GFP antibody conjugated to Alexa Fluor 488 (Abcam). All of the antibodies used were prepared and tested commercially and used at a low concentration to ensure specificity. Coverslips were mounted onto a slide to make a flow chamber, and freshly mixed imaging buffer (1 mg ml⁻¹ glucose oxidase, 0.02 mg ml⁻¹ catalase, 10% glucose and 100 mM mercaptoethylamine) was added to the sample chamber.

Imaging was performed on a custom-built setup based on a Leica DMI 300 inverse microscope. A 639-nm (UltraLaser, MRL-FN-639-800)

and 488-nm (OBIS) laser was collimated into the microscope objective (Zeiss, HCX PL APO 63X NA = 1.47 OIL CORR TIRF), with the illumination at the objective focus adjusted to approximately 1.5 and 0.8 kW cm⁻² for 639-nm and 488-nm lasers, respectively. Emitted photons were sequentially collected by a sCMOS camera (Prime 95B, Photometrics) at 33 Hz (30 ms per frame) for 2,000 frames. STORM reconstruction was performed using an in-house written MATLAB script²⁸. Each point spread function was fitted using maximum likelihood estimation (MLE). Note that the pixel-specific noise of each pixel was calibrated, characterized as a Gaussian distribution and convolved with the Poisson short-noise distribution for MLE fitting²⁹. Broad-spectrum fluorescent beads (diameter of around 100 nm; TetraSpec, Thermo Fisher Scientific) were imaged in both the Alexa Fluor 488 and 647 channels and the positions of each bead in the two different channels were mapped using a second polynomial mapping algorithm and used to align images from different illuminations^{28,29}. For computer code, see Code availability section below.

Stable-isotope tracing, lactate secretion, metabolite extraction and gas chromatography–mass spectrometry analysis

For stable-isotope labelling experiments, DMEM (D5030, Sigma) containing uniformly ¹³C-labelled glucose (¹³C₆-glucose; Cambridge Isotope Laboratories) and 10% dialysed FBS (Thermo Fisher Scientific) was added to cells before metabolite extraction and gas chromatography–mass spectrometry (GC–MS) analysis. Cellular metabolites were extracted after a rinse with cold 0.9% saline solution using a methanol/water/chloroform extraction³⁰. For analysis of lactate secretion, cells were cultured in DMEM with 10% dialysed FBS for 24 h, and 5 µl of initial and conditioned medium was extracted in 250 µl of 80% methanol:water containing 5 nmol ¹³C₃-lactate (Cambridge Isotope Laboratories). Conditioned medium was subjected to centrifugation at 1,000g for 15 min at 4 °C to remove cellular debris before extraction. The flux of lactate secretion (nmol per cell per h) was calculated by measuring the molar accumulation of lactate in conditioned medium divided by the viable cell density over the 24-h incubation³¹. After extraction, the aqueous phase was evaporated to dryness under vacuum by SpeedVac (Thermo Fisher Scientific), followed by methoxyamine (MOX)-tBDMS derivatization as previously described³². Derivatized samples were analysed by GC–MS using a DB-35MS column (30 mm × 0.25 mm i.d. × 0.25 µm) installed in an Agilent 7890B gas chromatograph interfaced with an Agilent 5977B mass spectrometer³¹, and corrected for natural isotope abundance using in-house algorithms adapted from previous work³³.

Extracellular acidification rate

ECAR was measured using a Seahorse XFe96 analyser (Agilent). Cells were seeded into Seahorse XFe96 plates at either 1.0 × 10⁴, 1.5 × 10⁴ or 2.0 × 10⁴ cells per well 24 h before measurements. The following day, the medium was exchanged with DMEM (Sigma-Aldrich) supplemented with 25 mM glucose and 2 mM glutamine and incubated for 30 min at 37 °C in an incubator without CO₂. Respiratory and glycolytic rates were measured in response to sequential injections of oligomycin (2 µM) and 2-D-deoxyglucose (50 mM). Cells were immediately lysed using 30 µl per well of Reagent A (Bio-Rad) after each experiment, and the protein level was quantified using the DC protein assay kit (Bio-Rad). Glycolytic rates of individual wells were normalized to milligrams of total protein, as quantified by a standard curve.

Xenografts

SUIT2 and A549 xenografts were established in 4–8-week-old male NCG mice (Charles River). Mice were injected with 0.5 × 10⁶ *KRAS4A*^{-/-} or *KRAS4A*^{+/-} cells into their contralateral flanks, and xenografts were grown for at most 6 weeks. Tumours were measured every 3–4 days with calipers. When tumours reached approximately 1 cm along their longest axis the mice were analysed for glucose uptake by ¹⁸F-FDG PET and computerized tomography (CT) imaging. After imaging, the mice

were killed and tumours were excised, weighed and fixed in formalin. All animal protocols were approved by the NYU School of Medicine Institutional Animal Care and Use Committee (IACUC). Tumours were generated in mice under the IACUC protocol IA16-00051, which allows tumours to grow to a volume of 2.5 cm³ or 2 cm in diameter. The tumour size did not exceed that allowed by the protocol in any of the mice.

Micro-PET imaging

All mice were subjected to ¹⁸F-FDG-PET scans under IACUC protocol IA17-00566 using a micro-PET/micro-X-ray CT (µPET/µCT) scanner (Inveon MM, Siemens Medical Solutions) equipped with a PET detector ring that comprised 16 detector blocks, each containing a row of four lutetium oxyorthosilicate detectors for a total of 64 detectors capable of 1.4-mm isotropic spatial resolution. The scan consisted of a 60-min whole-body µPET acquisition after the injection of ¹⁸F-FDG. This was followed by a 100-µm µCT scan after each µPET acquisition to assess the attenuation correction for the ¹⁸F-FDG datasets. The mice were fasted for at least 8 h before an intravenous injection of ¹⁸F-FDG (dose of 300–400 µCi diluted in physiological saline to a volume of 200 µl). The injection of ¹⁸F-FDG was administered using an infusion pump (PHD2000, Harvard Apparatus) at a rate of 60 µl min⁻¹. The radiotracer was administered to each mouse right after the start of the PET acquisition, which allowed the pharmacokinetics of ¹⁸F-FDG uptake to be captured in the whole body of the mouse. The hybrid scanner was equipped with a M2M BIOVET module, which was used to continuously monitor the vital signs of the mouse. The imaging scan consisted of first placing each mouse in an induction chamber using 3–5% isoflurane exposure for 2–3 min until the onset of anaesthesia. The mouse was subsequently positioned prone on the bed pallet over a recirculating warm-water blanket in which 1.0–1.5% isoflurane was administered by a nose cone throughout the scan. Mice were monitored continuously throughout the scanning session via a respiration pad and a rectal temperature probe as a feedback for temperature regulation of the whole stage. The body temperatures of the mice were maintained between 36 °C and 37 °C using a circulating water-heating pad.

PET data analysis

FDG-PET data were analysed using the Inveon Research Workplace software (Siemens Medical Solutions). Each PET dataset was reconstructed as a three-dimensional dataset co-registered to a 100-µm isotopic µCT dataset for attenuation correction and overlaid to provide anatomical context. Tumour uptake was calculated as the standard uptake value (SUV), a common metric in PET analysis to account for the variations introduced by the injected dose, the time of injection and the weight of the mouse. SUV was measured over a region of interest (ROI) that circumvented the tumour mass by thresholding each individual volume from the µCT dataset using Hounsfield units. SUV time–activity curves of each ROI were interpreted as a surrogate marker of glucose uptake.

Statistical analysis

All statistical analyses were performed with GraphPad Prism v.8.1.1. Differences between groups of two and more than two were assessed using a Student's *t*-test and two-way ANOVA with Tukey's post hoc analysis for multiple comparisons, respectively. *P* < 0.05 was the cut-off used for statistical significance.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The authors declare that the data supporting the findings of this study are available within the paper and its Supplementary Information.

Code availability

The code used for super-resolution imaging is available at GitHub (<https://github.com/yiny02/direct-Triple-Correlation-Algorithm>) and in the supplementary information of a previous study²⁹.

26. Peled, M. et al. Affinity purification mass spectrometry analysis of PD-1 uncovers SAP as a new checkpoint inhibitor. *Proc. Natl Acad. Sci. USA* **115**, E468–E477 (2018).
27. Zhou, M. & Philips, M. R. Nitrogen cavitation and differential centrifugation allows for monitoring the distribution of peripheral membrane proteins in cultured cells. *J. Vis. Exp.* **126**, e56037 (2017).
28. Yin, Y., Lee, W. T. C. & Rothenberg, E. Ultrafast data mining of molecular assemblies in multiplexed high-density super-resolution images. *Nat. Commun.* **10**, 119 (2019).
29. Huang, F. et al. Video-rate nanoscopy using sCMOS camera-specific single-molecule localization algorithms. *Nat. Methods* **10**, 653–658 (2013).
30. Metallo, C. M. et al. Reductive glutamine metabolism by IDH1 mediates lipogenesis under hypoxia. *Nature* **481**, 380–384 (2012).
31. Grassian, A. R. et al. IDH1 mutations alter citric acid cycle metabolism and increase dependence on oxidative mitochondrial metabolism. *Cancer Res.* **74**, 3317–3331 (2014).
32. Lewis, C. A. et al. Tracing compartmentalized NADPH metabolism in the cytosol and mitochondria of mammalian cells. *Mol. Cell* **55**, 253–263 (2014).
33. Fernandez, C. A., Des Rosiers, C., Previs, S. F., David, F. & Brunengraber, H. Correction of ¹³C mass isotopomer distributions for natural stable isotope abundance. *J. Mass Spectrom.* **31**, 255–262 (1996).

Acknowledgements We thank D. Esposito and W. Gillette for supplying highly purified recombinant KRAS proteins. This work was funded by the National Institutes of Health (R01CA163489 and R01CA116034 to M.R.P.; R01CA157490, R01CA188048, P01CA117969, R35CA232124 and R01GM095567 to A.C.K.; R35CA210018 and U01CA217864 to A.B.; R01CA214948 to R.P.; T32CA009161 to J.P.M.; and T32GM088118 to C.R.A.); the AACR Basic

Cancer Research Fellowship (grant number 15-40-01-MAHA to J.P.M.); the Charles H. Revson Senior Fellowship in Biomedical Science (to J.P.M.), American Cancer Society-New York Cancer Research Fund Postdoctoral Fellowship (grant number PF-18-215-01-TBG to S.J.P.), UCSF Pancreas Center and the Schwartz Family Foundation (to W.-C.C.); and the Lustgarten Foundation and SU2C (to A.C.K.). Proteomic analysis and in vivo imaging were performed in the Proteomics and DART Preclinical Imaging Cores, respectively, each partially funded by the NYU Laura and Isaac Perlmutter Cancer Center Support Grant, NIH/NCI P30CA016087. The Center for Advanced Imaging Innovation and Research (CAI2R, www.cai2r.net) at New York University School of Medicine is supported by NIH/NIBIB P41 EB017183.

Author contributions C.R.A., J.P.M. and M.R.P. designed and interpreted all experiments and wrote the manuscript. Unless otherwise stipulated, J.P.M. and C.R.A. performed all experiments. W.-C.C. and A.B. provided advice and A549 and SUIT2 cells that were engineered by CRISPR–Cas9 to lack KRAS4A. M.Z. performed mitochondrial purifications. I.M.A. and H.C. performed the 2-DG growth inhibition studies. J.S. performed hexokinase activity assays. S.L.M. performed the PET CT studies. S.J.P. performed the Seahorse analysis and ¹³C-glucose labelling. M.J.M. performed the super-resolution microscopy. E.G., A.C.K., Y.Z.W., R.P., S.R.H., and E.R. assisted with the interpretation of the results and edited the manuscript.

Competing interests A.C.K. has financial interests in Vescor Therapeutics. A.C.K. is an inventor on patents that pertain to KRAS-regulated metabolic pathways, redox control pathways in pancreatic cancer, targeting GOT1 as therapeutic approach and the autophagic control of iron metabolism. A.C.K. is on the Scientific Advisory Board of Rafael Pharmaceuticals and has been a consultant for Diciphera Pharmaceuticals.

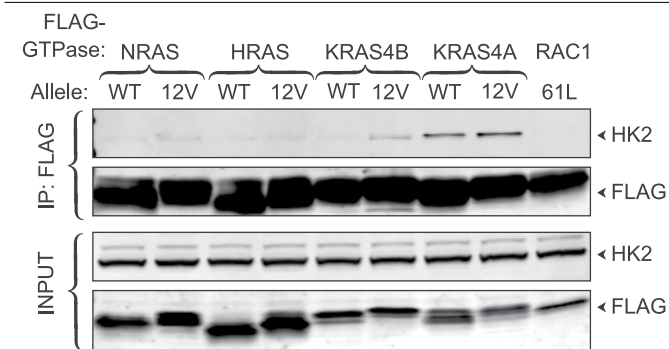
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1832-9>.

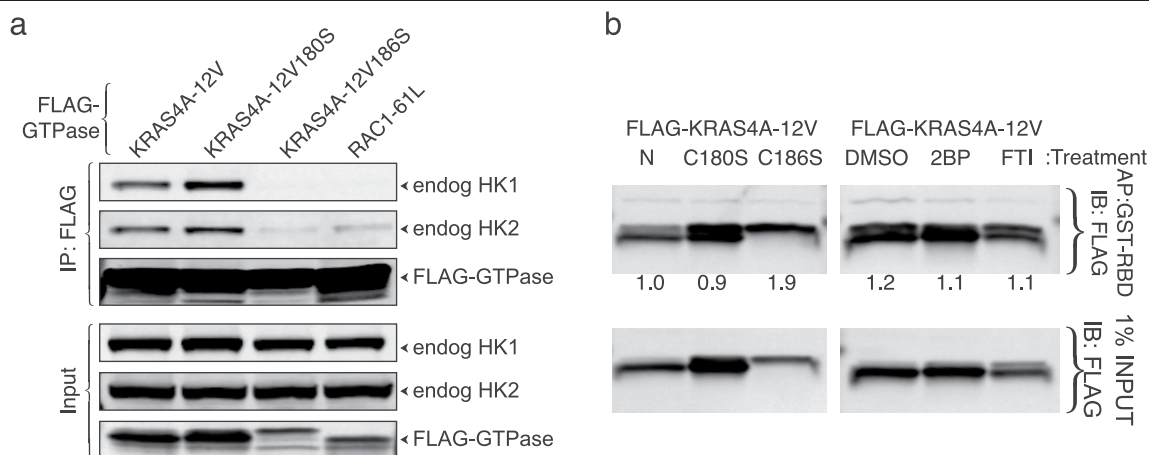
Correspondence and requests for materials should be addressed to M.R.P.

Peer review information *Nature* thanks Dimitrios Anastasiou and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

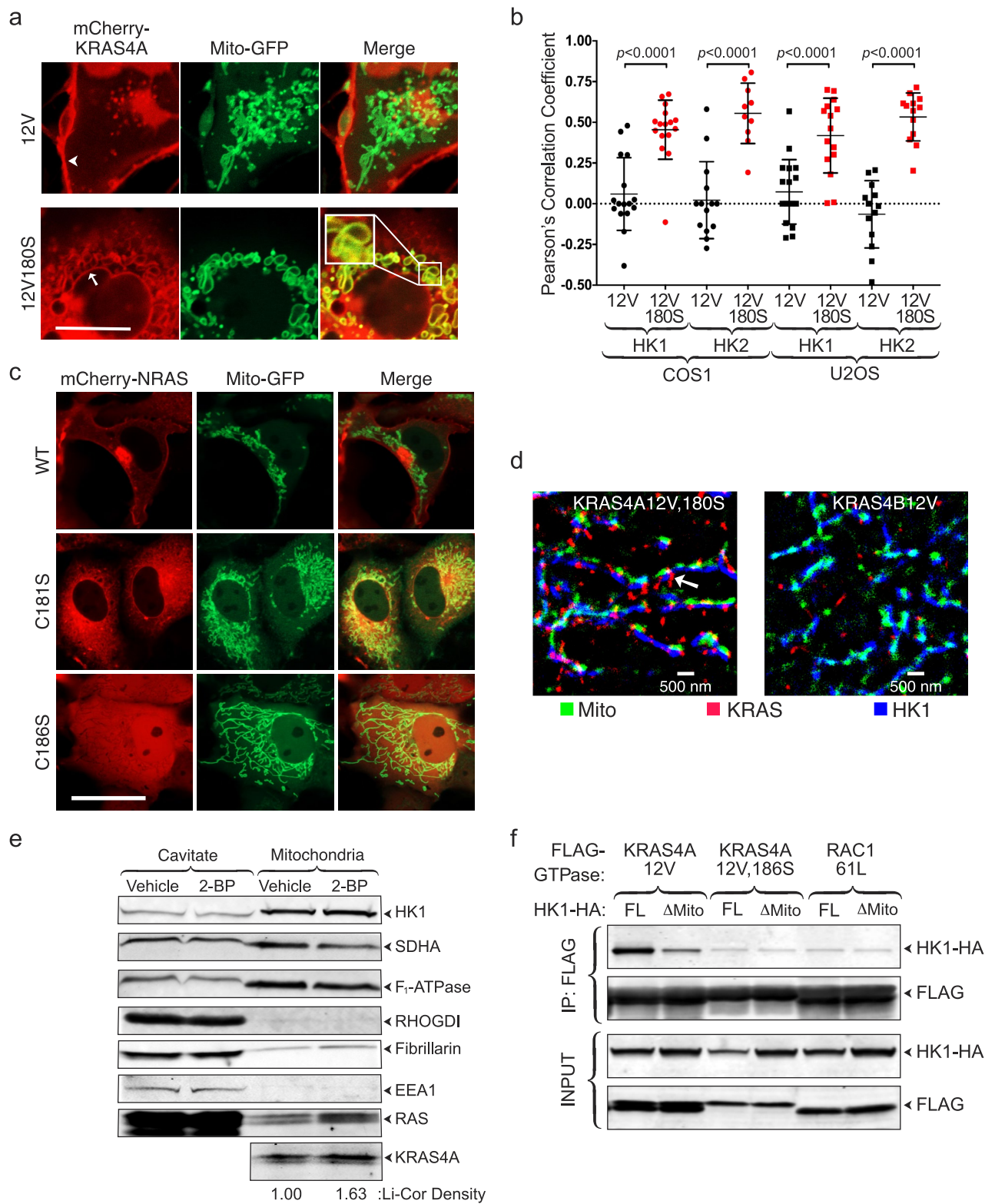


Extended Data Fig. 1 | KRAS4A binds to HK2 in an isoform-specific and GTP-dependent manner. The indicated Flag-tagged RAS constructs (with or without G12V mutations) were expressed in HeLa cells and immunoprecipitated with anti-Flag beads. Blots were probed for Flag-tagged proteins and endogenous HK2. Flag-RAC1(Q61L) served as the negative control. The immunoblot shown is representative of four independent experiments.



Extended Data Fig. 2 | Association of KRAS4A with HK1 and HK2 requires prenylation but is diminished by palmitoylation. a, HeLa cells expressing the indicated, Flag-tagged GTPase were lysed, KRAS4A or RAC1 were immunoprecipitated and the precipitates were blotted for Flag-tagged GTPases or endogenous HK1 or HK2. **b**, To confirm that the results in **a** reflect membrane targeting rather than GTP loading, the relative GTP loading of the KRAS4A proteins was determined by GST-RAF1-RBD affinity capture. Left, Flag-KRAS4A with an activating G12V mutation and either a native (N) membrane-targeting sequence or one with the indicated substitution were

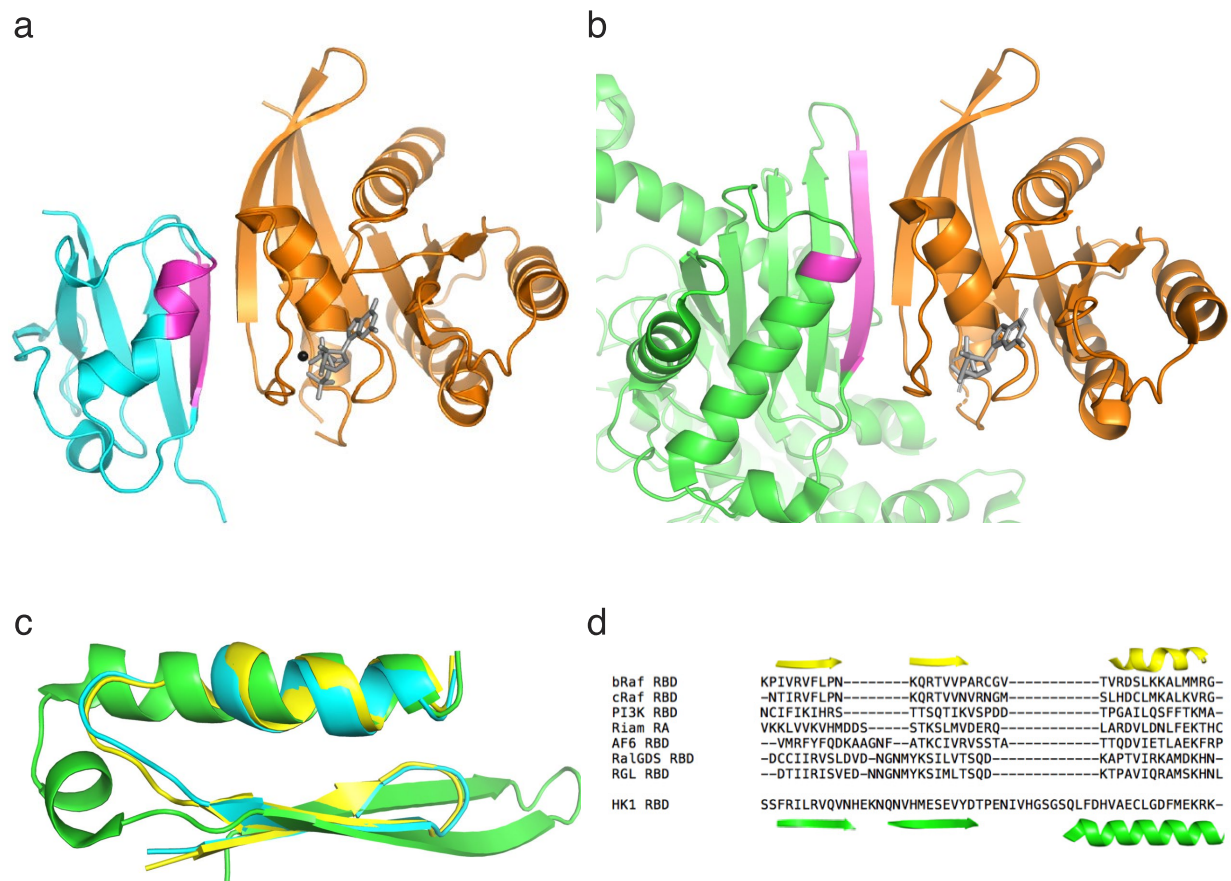
expressed in HEK293 cells. Right, in addition, cells expressing Flag-KRAS4A(G12V) were treated with 2-BP to inhibit palmitoylation or FTI to inhibit farnesylation. The total level of Flag-KRAS(G12V) was measured by anti-Flag immunoblot of 1% of the lysate (bottom) and GTP-bound Flag-KRAS4A was measured by affinity purification of the remaining lysate with GST-RAF1-RBD (top). The number under each lane is the amount of GTP-bound KRAS4A relative to lane 1 after normalization for expression (bottom). The immunoblots shown are representative of two independent experiments (**a**, **b**).



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Colocalization of palmitoylation-deficient KRAS4A but not NRAS with HK1 on mitochondria. **a**, Representative live-cell images of COS1 cells co-transfected with the indicated mCherry-KRAS4A constructs and GFP extended with the mitochondrial targeting sequence of HK1 (Mito-GFP). The arrowhead and the arrow indicate the plasma membrane and nuclear envelope, respectively. The cell shown is representative of hundreds on each plate of five independent transfections. Scale bar, 10 μ m. **b**, Colocalization of KRAS4A (with or without palmitoylation) with HK1 and HK2. COS1 or U2OS cells were co-transfected with GFP-tagged HK1 or HK2 and mCherry-tagged, constitutively active KRAS4A(G12V), with or without mutation of cysteine 180 to serine (C180S) to block palmitoylation. The cells were imaged alive using a Zeiss 800 laser scanning confocal microscope and the Pearson's correlation coefficient between the red and green channels was measured. Data are mean \pm s.d. of the values measured in $n = 15$ cells examined. Significance was determined by unpaired, two-tailed Student's *t*-test. **c**, Neither wild-type nor palmitoylation-deficient NRAS colocalizes with Mito-GFP on mitochondria. mCherry tagged wild-type, palmitoylation-deficient (C181S) or prenylation-deficient (C186S) NRAS were co-expressed in COS1 cells with Mito-GFP and imaged alive using an inverted Zeiss 800 laser scanning confocal microscope.

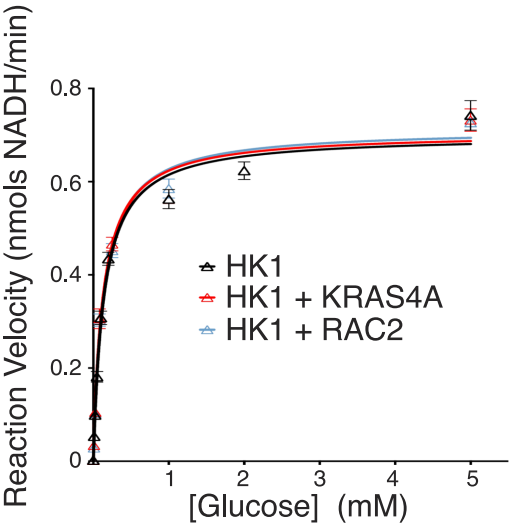
Scale bar, 10 μ m. The images shown are representative of hundreds of transfected cells on each plate in two independent experiments. **d**, Super-resolution (STORM) image of U2OS cells transfected with Flag-KRAS4A(G12V/C180S) or Flag-KRAS4B(G12V), showing colocalization with HK1 on the OMM (arrow) of KRAS4A(G12V/C180S) but not KRAS4B(G12V) ($n = 3$). Mito, mitochondria. **e**, Mitochondria were purified from HCT-15 cells that were pre-treated with vehicle or 2-BP and analysed by immunoblot with the indicated antibodies: succinate dehydrogenase (SDHA; mitochondrial matrix), F₁-ATPase (inner mitochondrial membrane), RHOGDI (cytosol), fibrillarin (nucleolus) or EEA1 (endosomes). RAS indicates the total level of RAS detected by a pan-RAS antibody. The KRAS4A immunoblot was quantified by a Li-Cor Odyssey infrared scanner. The immunoblot shown is representative of two independent experiments. **f**, The interaction of KRAS4A with HK1 requires the HK1 OMM-targeting sequence. The indicated Flag-tagged KRAS4A constructs were co-expressed in HEK293 cells with HA-tagged full-length HK1 (FL) or HK1 missing its OMM-targeting region ($\Delta 1-21$; Δ Mito). Flag-KRAS4A was immunoprecipitated, and binding to HK1 was assessed with an anti-HA immunoblot. The immunoblot shown is representative of four independent experiments.



Extended Data Fig. 4 | A putative RAS-binding region in HK1. a, Crystal structure (PDB 4G0N) of the CRAF RBD (cyan and magenta) in complex with the G domain of HRAS (orange; nucleotide (grey) and magnesium (black)). The regions of the CRAF RBD that mediate the interaction with HRAS are magenta. **b**, Superposition of the N-terminal lobe of HK1 (green and magenta; PDB 4F90) alongside HRAS (orange). The putative region of HK1 that interacts with RAS, corresponding to that of the CRAF RBD, is coloured magenta. **c**, A section of the

helix-loop-sheet structure that is common to RBDs is superimposed on the region of HK1 highlighted in **b**. The RBD of CRAF is shown in cyan, BRAF (PDB 3NYS) in yellow and the putative RBD of HK1 in green. **d**, Sequence alignments of validated RBDs and the HK1 putative RBD shown with structural motifs of BRAF (yellow) and HK1 (green). Despite the highly conserved structural features shown in **c**, RBDs have little sequence homology.

a



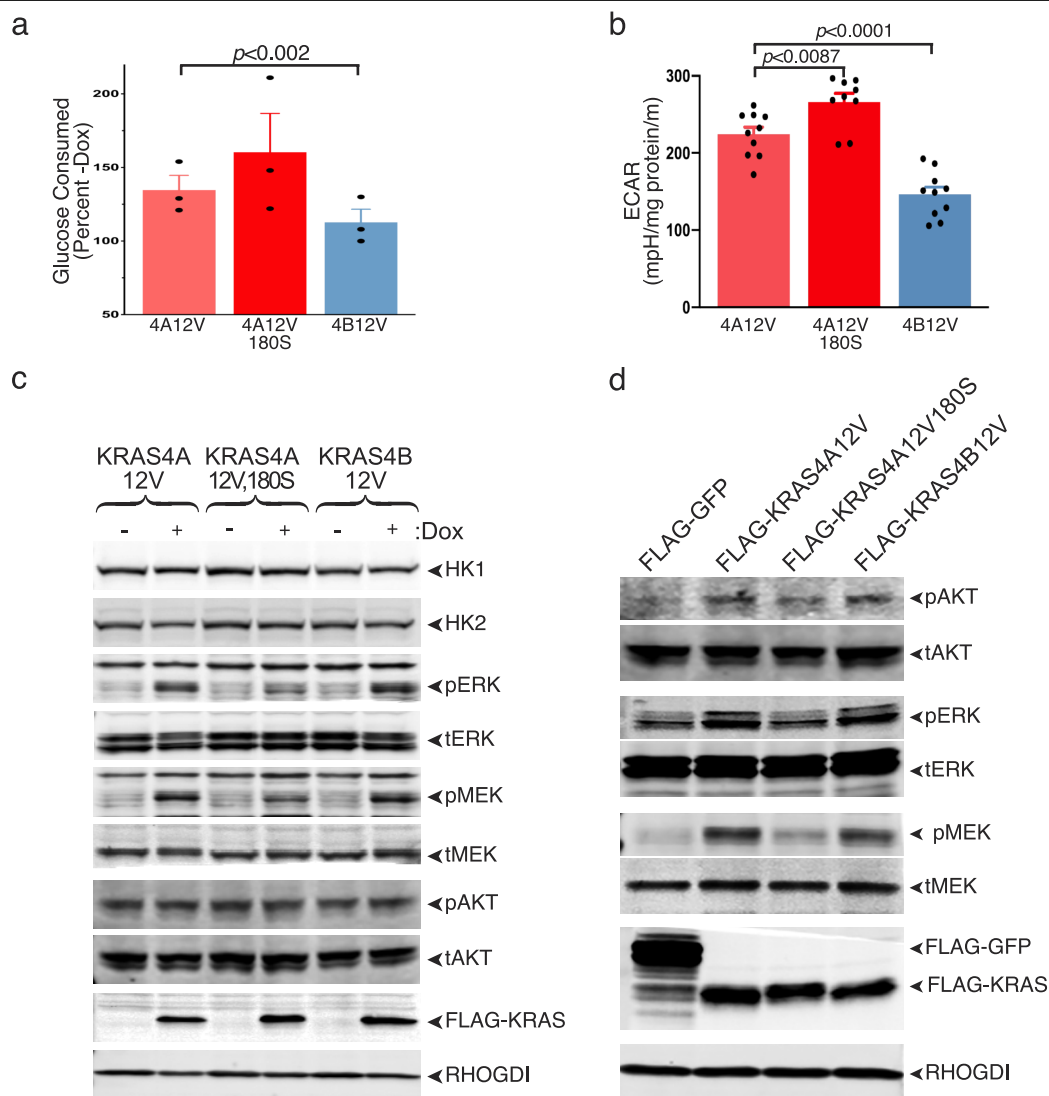
b

Extended Data Table 2. Hexokinase Enzyme Kinetics ± 2-Deoxyglucose (2DG) ± RAC2 or KRAS4A

	HK1 Full Length			HK1 C-term			HK2 Full Length		
	V_{MAX}	K_M	R^2	V_{MAX}	K_M	R^2	V_{MAX}	K_M	R^2
	(nmols/min)	(mM)		(nmols/min)	(mM)		(nmols/min)	(mM)	
Enzyme only	0.98±0.03	0.15±0.03	0.885	1.01±0.07	0.36±0.11	0.849	1.76±0.10	0.32±0.08	0.859
+ 2-DG	0.81±0.05	1.10±0.23	0.873	1.26±0.17	3.23±1.12	0.861	1.65±0.15	1.27±0.36	0.816
+ 2-DG + RAC2	0.80±0.09	1.41±0.49	0.729	1.34±0.28	3.98±2.02	0.771	1.93±0.27	2.63±1.00	0.803
+ 2-DG + KRAS4A	1.02±0.09	0.80±0.24	0.742	1.32±0.11	3.74±0.76	0.950	1.82±0.12	1.79±0.33	0.923

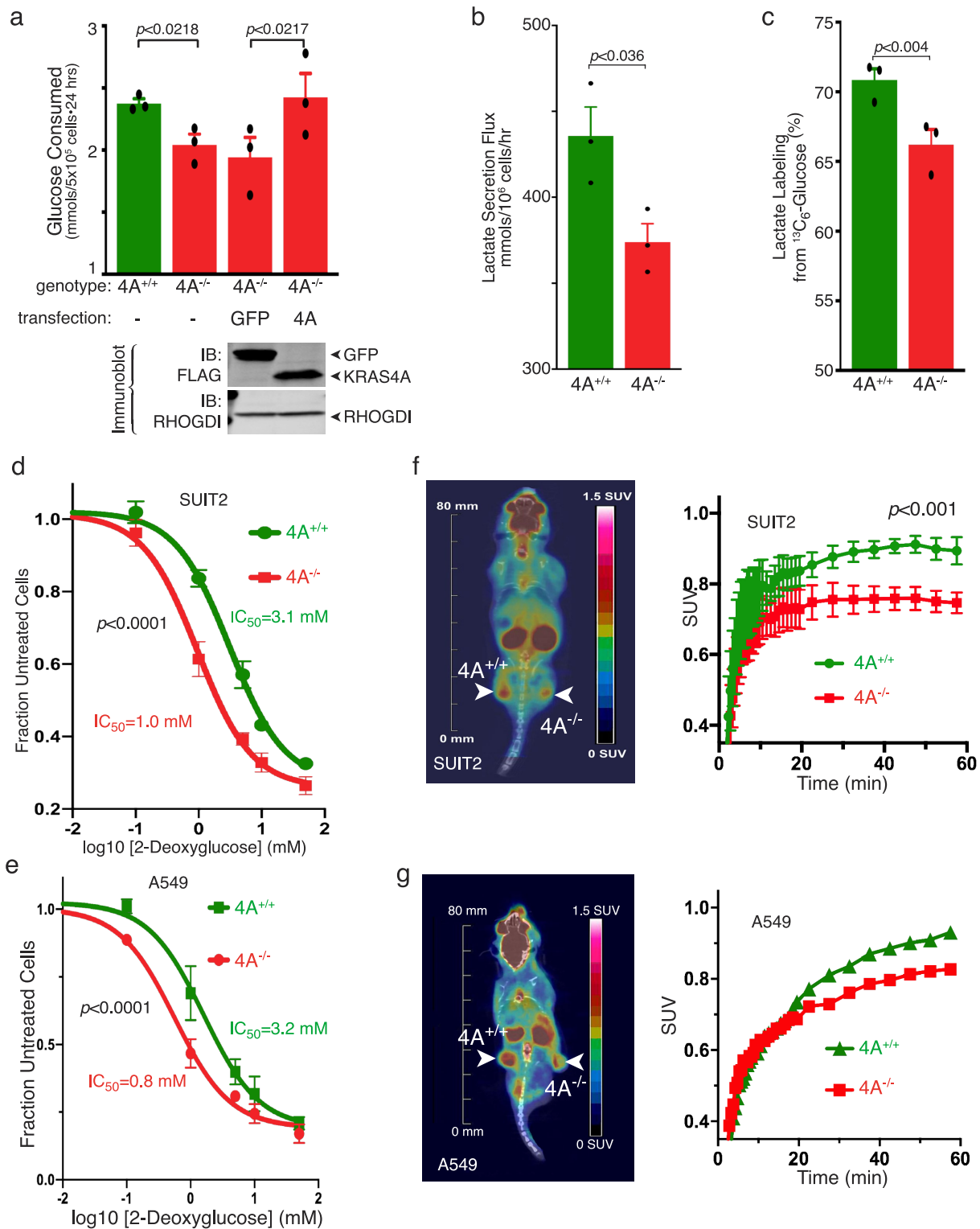
Extended Data Fig. 5 | Enzyme kinetics of hexokinases with and without recombinant KRAS4A and with and without 2-DG. **a**, The activity of recombinant full-length HK1 is unaffected by recombinant KRAS4A. Reaction velocity is plotted (mean ± s.e.m.) as a function of glucose concentration. Velocities are plotted with or without the addition of recombinant, GTP-loaded KRAS4A or RAC2. Plots combine independent assays ($n = 4$). **b**, Enzyme kinetics of hexokinases. Full-length HK1 and HK2 and the catalytic C-terminal domain of

HK1 were expressed in *E. coli* as GST fusion proteins and affinity purified with glutathione-agarose beads. Hexokinase activity on the beads was measured with a linked assay kit (BioVision) in which the glucose-6-phosphate produced is oxidized by glucose-6-phosphate dehydrogenase to form NADH, which reduces a colourless probe to a coloured product with strong absorbance at 450 nm. V_{max} and K_m were calculated by nonlinear regression using GraphPad Prism v 8.1.1 and the goodness of fit is given as R^2 .



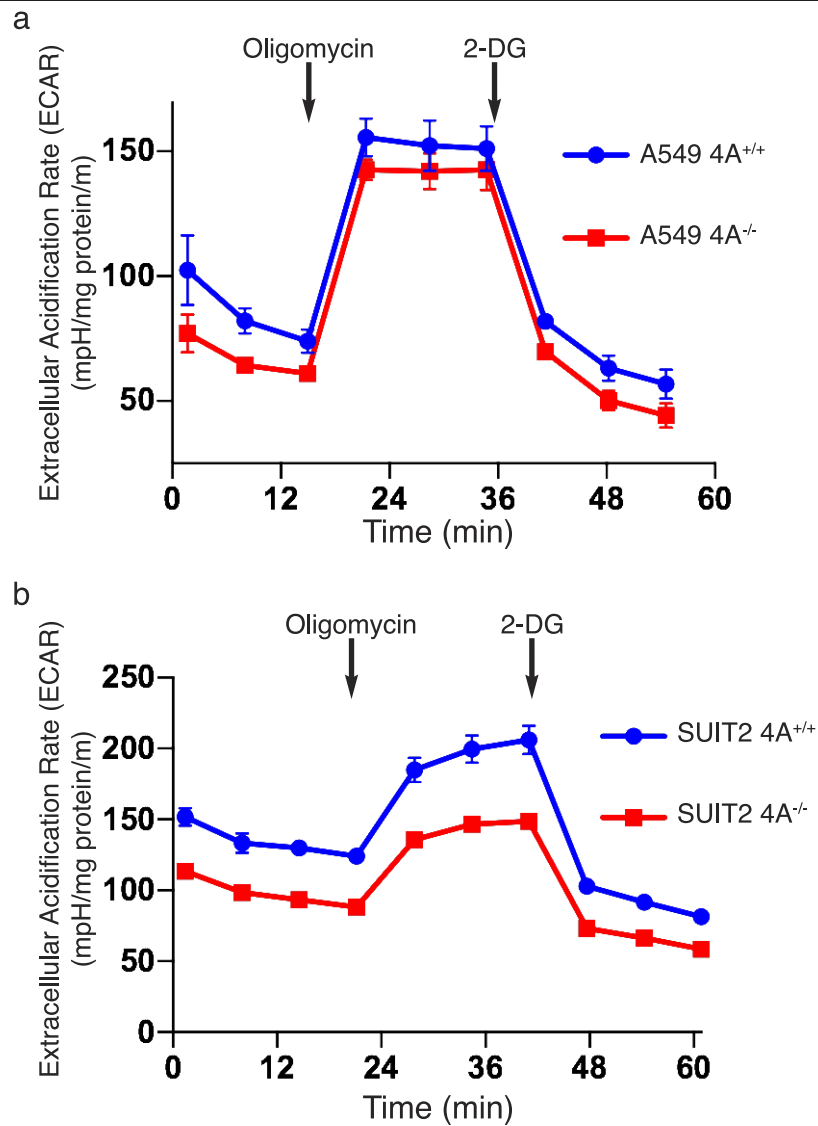
Extended Data Fig. 6 | Dissociation between MAPK signalling and differential stimulation of glucose consumption and basal ECAR by KRAS4A versus palmitoylation-deficient KRAS4A and KRAS4B. **a–c**, Flp-In T-REx 293 cells were generated that express the indicated KRAS proteins after induction with doxycycline. **a**, **b**, Glucose consumption (mean \pm s.e.m.; $n = 5$) (**a**) and basal ECAR (mean \pm s.e.m.; $n = 10$) (**b**) were measured in doxycycline-induced cells. This revealed the order of potency to be KRAS4A(G12V/C180S) > KRAS4A(G12V) > KRAS4B(G12V). Significance was determined by Student's *t*-test (paired in **a**; unpaired in **b**). **c**, Immunoblot reveals equivalent expression of the three KRAS proteins in the cells used in **a** and **b**. Whereas KRAS4A(G12V)

and KRAS4B(G12V) induced equivalent levels of phosphorylated (p)ERK and phosphorylated MEK, KRAS4A(G12V/C180S) (which is palmitoylation deficient) was less potent. These cells have constitutively high levels of AKT phosphorylation that were not altered by expression of any form of KRAS. Note also that despite MAPK stimulation, protein levels of HK1 and HK2 were not altered. The immunoblots shown are representative of two independent experiments. **t**, total. **d**, Parental HEK293 cells with lower basal levels of phosphorylated AKT were transfected with the indicated constructs, transferred to 0.1% serum 18 h after transfection and lysed 24 h later. Lysates were analysed for the indicated proteins by immunoblot ($n = 2$).



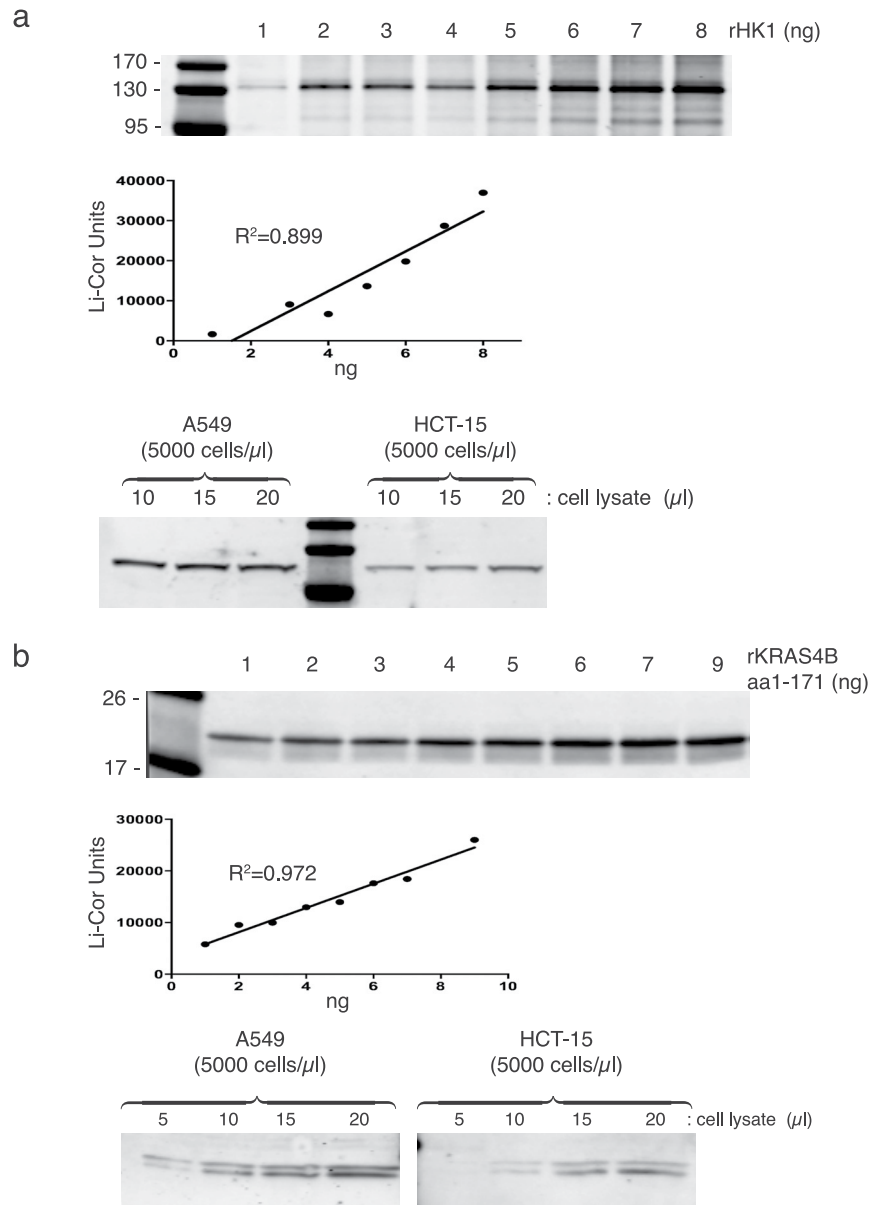
Extended Data Fig. 7 | Glucose consumption and 2-DG sensitivity in *KRAS*-mutant human tumour cells with or without exon 4A. **a**, Rate of glucose consumption (mean \pm s.e.m.; $n = 3$) in parental (4A^{+/+}) and exon 4A-targeted (4A^{-/-}) A549 cells transfected with Flag-GFP or Flag-KRAS4A(G12V). Representative expression is shown by immunoblot. **b**, Flux of lactate secretion (mean \pm s.e.m.; $n = 3$) measured over 24 h in A549 cells with the indicated genotype. **c**, Incorporation of ¹³C from glucose into lactate (mean \pm s.e.m.; $n = 3$) in A549 cells with the indicated genotype. Significance in **a–c** was determined by paired Student's *t*-test. **d**, **e**, Growth inhibition by 2-DG of SUIT2 (**d**; 48 h) and A549 (**e**; 24 h) cells with or without the 4A exon of *KRAS* (mean \pm s.e.m.; $n = 3$;

significance by two-way ANOVA). **f**, **g**, SUIT2 (**f**) or A549 (**g**) cells were used to establish xenograft tumours on the contralateral flanks (4A^{+/+} versus 4A^{-/-}) of NCG mice. Six weeks later, when the tumours were established and of equivalent size, glucose uptake was measured by ¹⁸F-FDG-PET CT scan. **f**, Left, coronal scan of a representative mouse with SUIT2 xenografts. Glucose uptake is represented by a colour look-up table. Right, SUVs of the entire tumour are plotted (mean \pm s.d.; $n = 5$ mice; significance by two-way ANOVA) as a function of time after ¹⁸FDG injection. **g**, Left, coronal scan of a representative mouse (of $n = 5$ mice) with A549 xenografts. Right, SUV versus time after injection or that mouse.



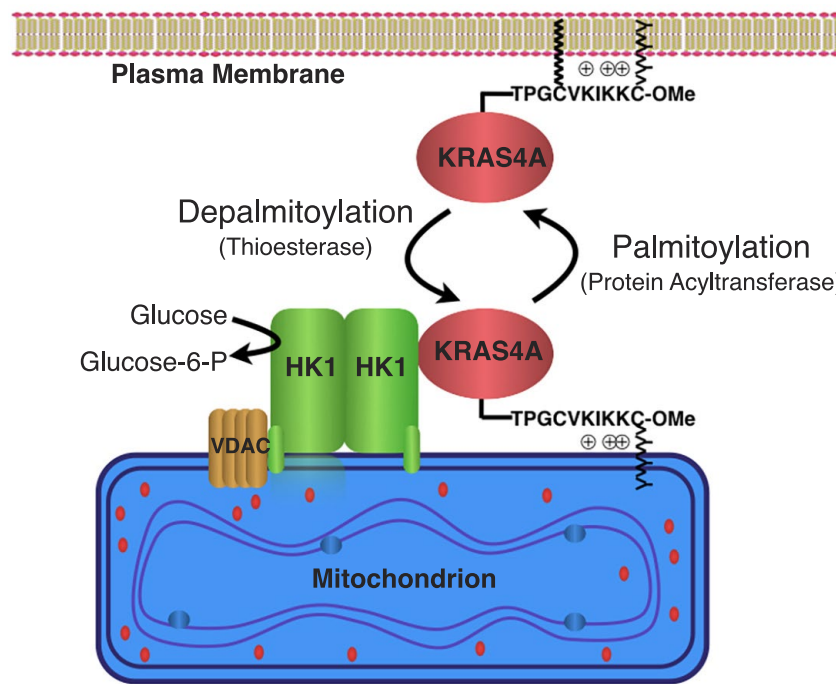
Extended Data Fig. 8 | Diminished basal ECAR in *KRAS*-mutant tumour cells in which *KRAS* exon 4A is disrupted. a, b, ECAR measured by Seahorse XFe96 of A549 (a) and SUI2 (b) cells with or without disruption of *KRAS* exon 4A by

CRISPR-Cas9. Oligomycin inhibits oxidative phosphorylation and allows the glycolytic reserve to be measured. 2-DG inhibits glycolysis. Data are mean \pm s.e.m. ($n=10$ technical replicates).



Extended Data Fig. 9 | Quantification of HK1 and total KRAS in tumour cell lines. **a**, Top, a standard curve for the quantification of HK1 by immunoblot was generated by titrating recombinant GST-HK1 (rHK1) and probing with a rabbit monoclonal antibody (Cell Signaling Technology, C3534). Bottom, immunoblots of the indicated amounts of lysate (5,000 cells per μ l). Calculations based on these results and a molecular mass of 100 kDa indicate 200,000 and 150,000 molecules per cell for A549 and HCT-15 cells, respectively. **b**, Top, a standard curve for the quantification of total KRAS by

immunoblot was generated by titrating recombinant KRAS4B truncated at amino acid 171 and probing with a mouse monoclonal antibody (Sigma-Aldrich, WH0003845M1). Bottom, immunoblots of the indicated amounts of lysate (5,000 cells per μ l). Calculations based on these results and a molecular mass of 21 kDa indicate 700,000 and 200,000 molecules per cell for A549 and HCT-15 cells, respectively. The standard curves in **a**, **b** were plotted by linear regression using GraphPad Prism v.8.1.1 and the goodness of fit is given as R^2 .



Extended Data Fig. 10 | Model of KRAS4A regulation of HK1. Like all palmitoylated GTPases, KRAS4A cycles between a palmitoylated and a depalmitoylated state. When palmitoylated, the protein has relatively high affinity for the plasma membrane, owing to farnesylation of the C-terminal CAAX sequence and an adjacent polybasic region that operates in conjunction with palmitoylation. After depalmitoylation, KRAS4A loses affinity for the

plasma membrane and gains affinity for endomembranes, including the OMM. Tethering of KRAS4A to the OMM allows it to interact with HK1, which is resident on this compartment owing to an N-terminal OMM-targeting sequence and protein–protein interaction with VDAC. Interaction of KRAS4A with HK1 on the OMM decreases allosteric inhibition by glucose-6-phosphate and thereby enhances HK1 activity and glycolytic flux.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Experimental STORM images were recorded using Micro-Manager (v 1.4). Confocal data were acquired with Zeiss Zen Blue 2 software version 2.1.

Data analysis

GraphPad Prism (v. 8.1.1) was used to test significance and determine enzyme kinetics. Proteome Discoverer (v2.1) was used to identify peptides in MS by searching the UniProt Database. Pearson's correlation coefficients of confocal images were calculated with Zeiss Zen Blue 2 version 2.1 software. STORM images were reconstructed using a custom written program performed using Matlab (v2017b), C++ (via Intel Core i7 7800X) and CUDA8.0 (via NVIDIA GTX 1060).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The authors declare that [the/all other] data supporting the findings of this study are available within the paper [and its supplementary information files].

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The data are almost entirely based on cultured human cells. No biological specimens were examined and reported such that sample size does not apply. Except where otherwise stated, all experiments on cultured cells were performed three or more times so as to allow tests of significance using the Student's t test. For the xenograft mouse imaging studies three mice were examined with tumors derived from SUIT2 cells and 5 mice were examined with tumors derived from A549 cells. Each mouse served as its own control since tumors of each genotype were injected into contralateral flanks. This allowed for reduced sample size. Because the pilot study showed a marked difference in 18FDG uptake between the two genotypes, and because of the high cost associated with animal PET imaging, no sample size calculation was performed but instead we proceeded to test each cell type (with and without KRAS4A) in five mice for each of the two cell types. Tail vein injection failed in one of the SUIT2 mice and another died during imaging. All 8 mice successfully injected and scanned gave the same result (greater uptake in KRAS4A expressing tumors at all times after injection), making the data highly significant by two-way ANOVA.
Data exclusions	No data were excluded.
Replication	All experiments were repeated at least twice and the vast majority were repeated 3-5 times to assure reproducibility. Although the magnitude of the differences varied among replications, in no replicate was there no difference or a reversal of rank order.
Randomization	No assignment to experimental groups pertain to these studies.
Blinding	The molecular biological, cell biological and biochemical analyses of cultured cells that make up the bulk of this study involved multiple conditions, genes or proteins that required multiple steps to process including plasmid cloning, protein and plasmid purification, transfection, culture and analysis by SDS-PAGE or microscopy. A single scientist performs each of the multiple steps and must keep careful track of conditions. It would be exceedingly difficult to blind such studies. Blinding is not standard operating procedure in this type of work. The 18FDG PET/CT analysis of mice was performed in a blinded fashion.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used

With the exception of the polyclonal rabbit anti-human KRAS4A antibody that we developed and licensed to Millipore for commercial distribution (now available as Sigma-Aldrich ABC1442), all antibodies were purchased from the sources cited in the manuscript. Source, catalog number and dilutions used are: anti-FLAG (Sigma-Aldrich, F7425, Lot:085M4774V 1:2000), anti-GFP (ThermoFisher, A-6455, Lot:1826342, 1:2000), anti-HK1 (CST, #2024, Clone:C35C4, Lot:3 1:2000), anti-HK2 (CST, #2867, Clone:C64G5, Lot:3, 1:2000), anti-SDHA (CST, #11998, Clone:D6J9M, Lot:2, 1:1000), anti-pMEK (CST, #9121, Lot:47, 1:1000), anti-tMEK (CST, #4694, Clone:L38C12, 1:1000), anti-pERK (CST, #9106, Clone:E10, 1:1000), anti-pAKT (CST, #9271, Clone:S473, Lot:14, 1:1000), anti-tAKT (CST, #9272, Lot:27, 1:1000), anti-pan-RAS (CalBiochem, OP40, Clone:Ab-3, Lot: D00119097, 1:2000), anti-Fibrillarin (Santa Cruz Biotechnology, sc-374022, Clone:G-8 Lot:GO116, 1:1000), anti-EEA1 (Santa Cruz Biotechnology, sc-137130, Clone:G-4, Lot:F2716, 1:1000), anti-F1-ATPase (Santa Cruz Biotechnology, sc-514419, Clone:C-12, Lot:E1016, 1:1000), anti-tERK (Santa Cruz Biotechnology, sc-94, K-23, 1:1000), anti-Rho GDI (Santa Cruz Biotechnology, sc-360, Clone:A-20, Lot:J0313, 1:5000), anti-HA (Santa Cruz Biotechnology, sc-7392, Clone: F-7, Lot:K3012 1:1000), anti-RAC2 (Abcam, 130415, Lot:GR310183-1, 1:1000), and anti-KRAS4A, polyclonal rabbit antibody developed by our lab available from Sigma-Aldrich as ABC1442 {Tsai, 2015 #4767}, 1:500.

Validation

All anti-RAS antibodies were validated by loss of immunoreactive band on Western blot analysis upon silencing RAS with siRNA or CRISPR Cas9. We recently published an exhaustive analysis of commercially available anti-RAS antibodies (see PMID 28951536).

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)

HEK293, U2OS, COS1, A549, SUIT2, Hct-15, Hct-116, HeLa were purchased from ATCC. Flp-In T-REx293™ were purchased from Fisher Scientific.

Authentication

Authentication was performed by the vendors. Each cell line was validated by short tandem repeat profiling.. In the case of Flp-In T-REx293™ the ability to insert cDNAs upon expression of Frp flipase also validated the line.

Mycoplasma contamination

All cell lines were checked for mycoplasma with the MycoAlert™ kit from Lonza. Any cell line found to be contaminated was cured with BM cyclin (Sigma-Aldrich; Cat. No. 10 799 050 001). All cells used in the experiments tested negative for mycoplasma.

Commonly misidentified lines
(See [ICLAC](#) register)

None.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals

NCG immunocompromised male mice were purchased from Charles River (Wilmington, MA, USA) and used for xenograft studies. They were 6-8 wks old at the time of tumor cell inoculation.

Wild animals

None

Field-collected samples

None

Ethics oversight

Laboratory animals housed within the New York University School of Medicine animal facilities were maintained in accordance with the Animal Welfare Act, the United States Department of Agriculture Regulations (9 CFR, Parts 1, 2, and 3), and the Guide for the Care and Use of Laboratory Animals (National Academy Press, Revised 1996). NYU Medical Center Animal Care & Use Program is fully accredited by the Association For Assessment And Accreditation Of Laboratory Animal Care International (AAALAC). New York University School of Medicine has a currently approved Animal Welfare Assurance Agreement (No. A3435-01) with the NIH Office for Protection from Research Risks. All animal care facilities and protocols are reviewed and approved by the NYU School of Medicine IACUC.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Multi-omics profiling of mouse gastrulation at single-cell resolution

<https://doi.org/10.1038/s41586-019-1825-8>

Received: 18 October 2018

Accepted: 22 October 2019

Published online: 11 December 2019

Ricard Argelaguet^{1,17}, Stephen J. Clark^{2,17*}, Hisham Mohammed^{2,17}, L. Carine Stapel^{2,17}, Christel Krueger², Chantiriolnt-Andreas Kapourani^{3,4}, Ivan Imaz-Rosshandler^{5,6}, Tim Lohoff^{2,5}, Yunlong Xiang^{7,8}, Courtney W. Hanna^{2,9}, Sebastien Smallwood², Ximena Ibarra-Soria¹⁰, Florian Buettner¹¹, Guido Sanguinetti³, Wei Xie^{7,8}, Felix Krueger¹², Berthold Göttgens^{5,6}, Peter J. Rugg-Gunn^{2,5,6,9}, Gavin Kelsey^{2,9}, Wendy Dean¹³, Jennifer Nichols⁵, Oliver Stegle^{1,14,15*}, John C. Marioni^{1,10,16*} & Wolf Reik^{2,9,16*}

Formation of the three primary germ layers during gastrulation is an essential step in the establishment of the vertebrate body plan and is associated with major transcriptional changes^{1–5}. Global epigenetic reprogramming accompanies these changes^{6–8}, but the role of the epigenome in regulating early cell-fate choice remains unresolved, and the coordination between different molecular layers is unclear. Here we describe a single-cell multi-omics map of chromatin accessibility, DNA methylation and RNA expression during the onset of gastrulation in mouse embryos. The initial exit from pluripotency coincides with the establishment of a global repressive epigenetic landscape, followed by the emergence of lineage-specific epigenetic patterns during gastrulation. Notably, cells committed to mesoderm and endoderm undergo widespread coordinated epigenetic rearrangements at enhancer marks, driven by ten-eleven translocation (TET)-mediated demethylation and a concomitant increase of accessibility. By contrast, the methylation and accessibility landscape of ectodermal cells is already established in the early epiblast. Hence, regulatory elements associated with each germ layer are either epigenetically primed or remodelled before cell-fate decisions, providing the molecular framework for a hierarchical emergence of the primary germ layers.

Recent technological advances have enabled the profiling of multiple molecular layers at single-cell resolution^{9–13}, providing novel opportunities to study the relationship between the transcriptome and epigenome during cell-fate decisions. We applied single-cell nucleosome, methylome and transcriptome sequencing¹² (scNMT-seq) to profile 1,105 single cells isolated from mouse embryos at four developmental stages (embryonic day (E)4.5, E5.5, E6.5 and E7.5) representing the exit from pluripotency and primary germ-layer specification (Fig. 1a–d, Extended Data Fig. 1). Cells were assigned to a specific lineage by mapping their RNA-expression profiles to a comprehensive single-cell atlas⁴ from the same stages when available or using marker genes (Extended Data Fig. 2). Using dimensionality reduction, we show that all three molecular layers contain sufficient information to separate cells by stage (Fig. 1b–d) and lineage identity (Extended Data Figs. 2, 3).

Epigenome dynamics at pluripotency exit

We characterized the changes in DNA methylation and chromatin accessibility during each stage transition. Globally, methylation levels increase from approximately 25% to approximately 75% in embryonic tissues and to about 50% in extra-embryonic tissues, driven mainly by a wave of de novo methylation from E4.5 to E5.5 that preferentially targets CpG-poor genomic loci^{6,8,14} (Fig. 1e, Extended Data Fig. 3). By contrast, we observed a more gradual decline in global chromatin accessibility from around 38% at E4.5 to around 30% at E7.5 (Fig. 1f), with no differences between embryonic and extra-embryonic tissues (Extended Data Fig. 3). To relate epigenetic changes to the transcriptional dynamics across stages, we calculated—for each gene and across all embryonic cells—the correlation between RNA expression and the corresponding DNA methylation or chromatin accessibility at the promoter. Out of

¹European Bioinformatics Institute (EMBL-EBI), Cambridge, UK. ²Epigenetics Programme, Babraham Institute, Cambridge, UK. ³School of Informatics, University of Edinburgh, Edinburgh, UK.

⁴MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. ⁵Wellcome–MRC Cambridge Stem Cell Institute, Jeffrey Cheah Biomedical Centre, University of Cambridge, Cambridge, UK. ⁶Department of Haematology, Jeffrey Cheah Biomedical Centre, University of Cambridge, Cambridge, UK. ⁷Center for Stem Cell Biology and Regenerative

Medicine, MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing, China. ⁸THU-PKU Center for Life Sciences, Tsinghua University, Beijing, China. ⁹Centre for

Trophoblast Research, University of Cambridge, Cambridge, UK. ¹⁰Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. ¹¹Helmholtz Zentrum München–German

Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany. ¹²Bioinformatics Group, Babraham Institute, Cambridge, UK. ¹³Department of

Biochemistry and Molecular Biology, Alberta Children's Hospital Research Institute, University of Calgary, Calgary, Alberta, Canada. ¹⁴European Molecular Biology Laboratory (EMBL),

Heidelberg, Germany. ¹⁵Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany. ¹⁶Wellcome Sanger Institute, Cambridge,

UK. ¹⁷These authors contributed equally: Ricard Argelaguet, Stephen J. Clark, Hisham Mohammed, L. Carine Stapel. *e-mail: stephen.clark@babraham.ac.uk; o.stegle@dkfz.de; john.marioni@

cruk.cam.ac.uk; wolf.reik@babraham.ac.uk

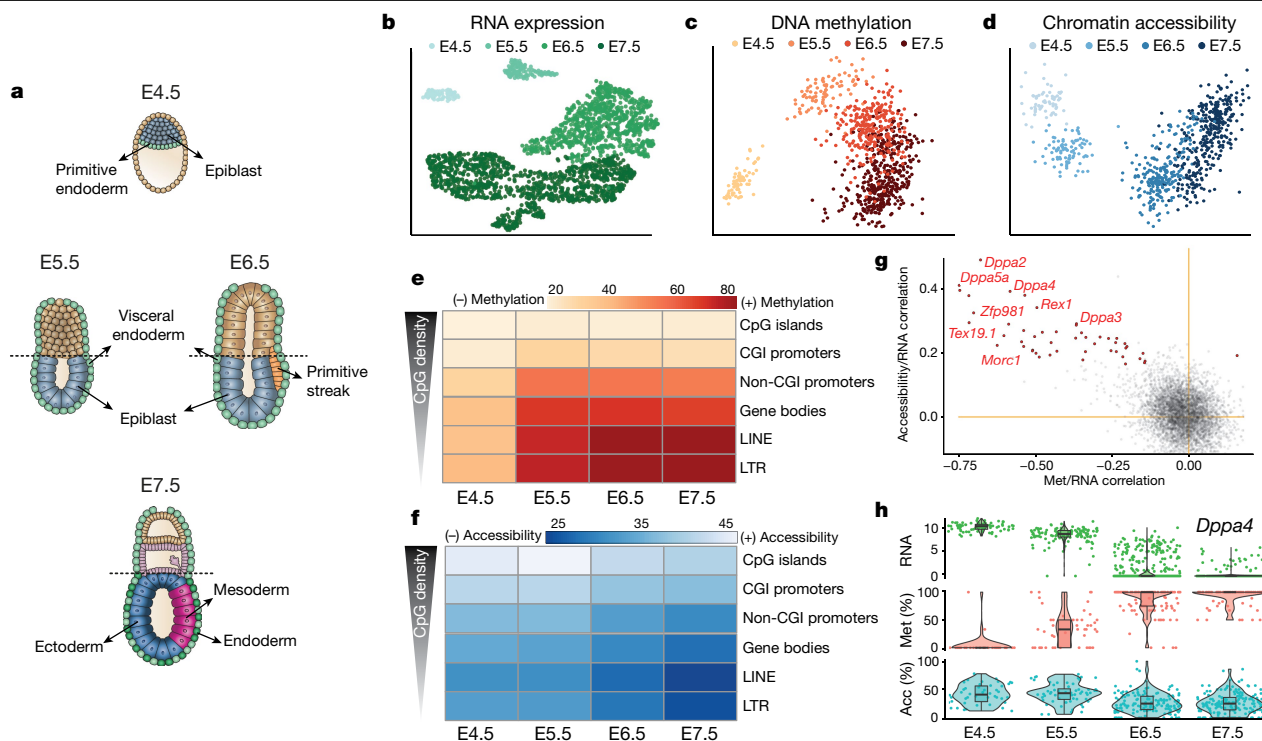


Fig. 1 | Single-cell multi-omics profiling of mouse gastrulation. **a**, Schematic of the developing mouse embryo, with stages and lineages considered in this study labelled. **b**, Dimensionality reduction of RNA-expression data using UMAP. Cells are coloured by stage. There are 1,061 cells included from 28 embryos sequenced using scNMT-seq and 1,419 cells from 26 embryos sequenced using scRNA-seq. **c**, **d**, Dimensionality reduction of DNA methylation data (**c**) and chromatin accessibility data (**d**) from scNMT-seq using factor analysis (Methods). Cells are coloured by stage. There are 986 cells included for DNA methylation data and 864 cells for chromatin accessibility data. **e**, **f**, Heat map of per cent DNA methylation levels (**e**) and per cent chromatin accessibility levels (**f**) by stage and genomic context. **g**, Scatter

plot of Pearson correlation coefficients of promoter methylation (Met) versus RNA expression (x axis) and promoter accessibility versus RNA expression (y axis). Each dot corresponds to one gene ($n = 4,927$). Red dots depict significant associations for both correlation types ($n = 39$, false discovery rate (FDR) $< 10\%$). Examples of early pluripotency and germ cell markers among the significant hits are labelled. **h**, Illustrative example of epigenetic repression of *Dppa4*. Box and violin plots show the distribution of RNA expression (log normalized counts, green), promoter methylation (red) and accessibility (Acc) (blue) per stage. Box plots show median levels and the first and third quartile, whiskers show $1.5 \times$ the interquartile range. Each dot corresponds to one cell.

5,000 genes tested, we identified 125 genes the expression of which shows significant correlation with promoter DNA methylation and 52 with expression significantly correlated with chromatin accessibility (Fig. 1g, Extended Data Fig. 4, Supplementary Tables 1, 2). These loci largely comprise markers of early pluripotency and germ cells, such as *Dppa4*, *Zfp42*, *Tex19.1* and *Pou3f1* (Fig. 1g, h, Extended Data Fig. 4), which are repressed, coinciding with the global increase in methylation and decrease in accessibility. In addition, this analysis identified genes, including *Trap1a* and *Zfp981*, that may have unknown roles in development. Notably, of the genes that are upregulated after E4.5, only 39 and 9 show a significant correlation between RNA expression and promoter methylation or accessibility, respectively (Extended Data Fig. 4). This suggests that the upregulation of these genes is probably controlled by other regulatory elements.

Characterizing germ-layer epigenomes

To understand the relationships between all three molecular layers during germ-layer commitment we next applied multi-omics factor analysis (MOFA)¹⁵ to cells collected at E7.5. MOFA performs unsupervised dimensionality reduction simultaneously across multiple data modalities, thereby capturing the global sources of cell-to-cell variability via a small number of inferred factors. Notably, the model makes use of multimodal measurements from the same cells, thereby detecting coordinated changes between the different data modalities.

As input to the model we used RNA-sequencing (RNA-seq) data across protein-coding genes and DNA methylation and chromatin accessibility

data across putative regulatory elements. This includes promoters and germ-layer-specific chromatin immunoprecipitation with DNA sequencing (ChIP-seq) peaks for distal H3K27ac (enhancers) and H3K4me3 (transcription start sites)¹⁶ (Extended Data Fig. 5). MOFA identified six factors, with the top two (sorted by variance explained) capturing the emergence of the three germ layers (Fig. 2a, b). Notably, MOFA links the variation at the gene-expression level to concerted methylation and accessibility changes at lineage-specific enhancer marks (Fig. 2c). By contrast, epigenetic changes at promoters or at H3K4me3-marked regions show much weaker associations with germ-layer formation (Fig. 2a, Extended Data Fig. 6, Supplementary Tables 3, 4). This supports other studies that have identified distal enhancers as lineage-driving regulatory regions^{8,17–19}. Inspection of gene–enhancer associations identified enhancers linked to key germ-layer markers including *Lefty2* and *Mesp2* (mesoderm), *Foxa2* and *Bmp2* (endoderm), and *Bcl11a* and *Sp8* (ectoderm) (Fig. 2c, Extended Data Fig. 7). Notably, ectoderm-specific enhancers display fewer associations than their mesoderm and endoderm counterparts, a finding that is explored further below.

The four remaining factors correspond to additional transcriptional and epigenetic signatures related to anterior–posterior axial patterning (factor 3), notochord formation (factor 4), mesoderm patterning (factor 5) and cell cycle (factor 6) (Extended Data Fig. 8).

Finally, we sought to identify transcription factors that could drive or respond to epigenetic changes in germ-layer commitment. Integrating differential-expression information with motif enrichment at differentially accessible loci revealed that lineage-specific enhancers were

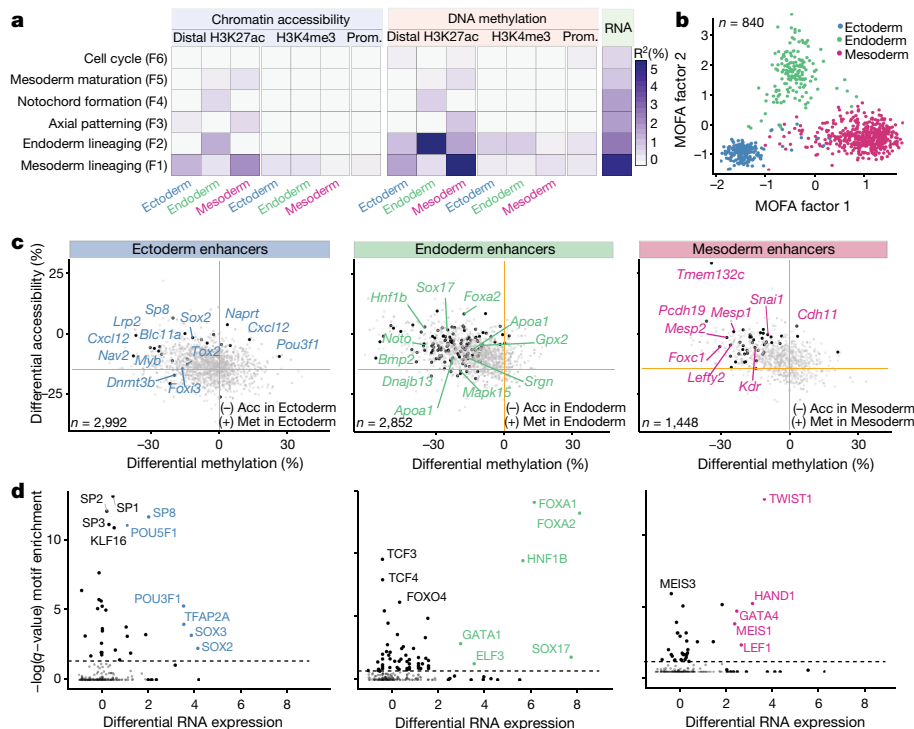


Fig. 2 | Multi-omics factor analysis reveals coordinated epigenetic and transcriptomic variation at enhancer elements during germ-layer commitment. **a**, Percentage of variance explained (R^2) by each MOFA factor (rows) across data modalities (columns). **b**, Scatter plot of MOFA factor 1 (x-axis) and MOFA factor 2 (y-axis). Cells are coloured according to their lineage assignment ($n = 840$). **c**, Scatter plots showing differential DNA methylation (x-axis) and chromatin accessibility (y-axis) at lineage-specific enhancers at E7.5. Ectoderm versus non-ectoderm cells (left, $n = 2,992$), endoderm versus non-endoderm cells (middle, $n = 2,852$) and mesoderm versus non-mesoderm cells (right, $n = 1,448$) are shown. Black dots depict gene–enhancer pairs with

significant changes in RNA expression and methylation or accessibility (Pearson's χ^2 test, $FDR < 10\%$). **d**, Transcription factor motif enrichment at lineage-defining enhancers. Motif enrichment (Fisher's exact test, $-\log(q \text{ value})$, y-axis, $n = 746$ motifs) versus differential RNA expression (log fold change, x-axis) of the corresponding transcription factor is shown. The analysis is performed separately for ectoderm- (left), endoderm- (middle) and mesoderm- (right) defining enhancers. Transcription factors with significant motif enrichment ($FDR < 1\%$) and differential RNA expression (edgeR quasi-likelihood test, $FDR < 1\%$) are labelled.

enriched for binding sites associated with key developmental transcription factors, including POU3F1, SOX2 and SP8 for ectoderm, SOX17, HNF1B, and FOXA2 for endoderm, and GATA4, HAND1 and TWIST1 for mesoderm (Fig. 2d).

Time resolution of the enhancer epigenome

We next investigated how the epigenomic patterns associated with germ-layer specification arise during development. DNA methylation levels in endoderm- and mesoderm-defining enhancers follow the genome-wide dynamics, increasing from an average of 25% to 80% in all cell types (Fig. 3, Extended Data Fig. 9). Upon lineage specification, they undergo concerted demethylation to about 50% in a cell-type-specific manner. The opposite pattern is observed for chromatin accessibility; accessibility of mesoderm- and endoderm-defining enhancers initially decreases from approximately 40% to 30% (following the genome-wide dynamics) before becoming more accessible (approximately 45%) upon lineage specification. The general dynamics of demethylation and chromatin opening of enhancers during embryogenesis are therefore apparently conserved in zebrafish, *Xenopus* and mouse¹⁹. Consistent with these data, when quantifying the H3K27ac levels of lineage-defining enhancers in more-differentiated tissues (E10.5 midbrain, E12.5 intestine and E10.5 heart)^{20,21}, we observe that a substantial number of enhancers remain marked by H3K27ac (Extended Data Fig. 5). This indicates that the enhancers established at E7.5 are, to a large extent, maintained later in development.

In contrast to the mesoderm and endoderm enhancers, the ectoderm enhancers are open and demethylated as early as E4.5 in the epiblast (Fig. 3, Extended Data Fig. 9). Only in cells committed to mesoderm and

fate do the ectoderm enhancers become partially repressed. Consistently, when measuring the accessibility dynamics at sites containing motifs for ectoderm-defining transcription factors (SOX2 and SP8), we find that these motifs are already accessible in the epiblast and lose accessibility specifically upon mesoderm commitment. Conversely, motifs associated with endoderm- and mesoderm-defining transcription factors become accessible in their respective lineages only at E7.5 (Extended Data Fig. 9).

These observations can be explained by either priming of an ectodermal signature in the epiblast or the maintenance of a pluripotency signature in the ectoderm. To investigate this, we overlapped the E7.5 enhancer annotations with published H3K27ac ChIP-seq data from embryonic stem cells (ES cells) and E10.5 midbrain^{21,22}. The E7.5 ectoderm enhancers display almost-exclusively pluripotent or neural signatures with notably different DNA methylation and chromatin accessibility dynamics (Extended Data Fig. 10). Pluripotency enhancers show an increase in methylation and a decrease in accessibility over time, suggesting a repression of these enhancers with similar dynamics to promoters of pluripotency genes (Fig. 1g, h). By contrast, neuroectoderm enhancers remain hypomethylated and accessible from E4.5 (Extended Data Fig. 10).

Finally, to infer temporal dependencies of enhancer activation, we used the RNA-expression profiles to order cells across two trajectories corresponding to mesoderm and endoderm commitment (Extended Data Fig. 11). By plotting the average DNA methylation and chromatin accessibility for each class of lineage-defining enhancer, we find that the methylation gain (and accessibility loss) of ectoderm enhancers precedes the demethylation (and accessibility gain) of mesoderm and

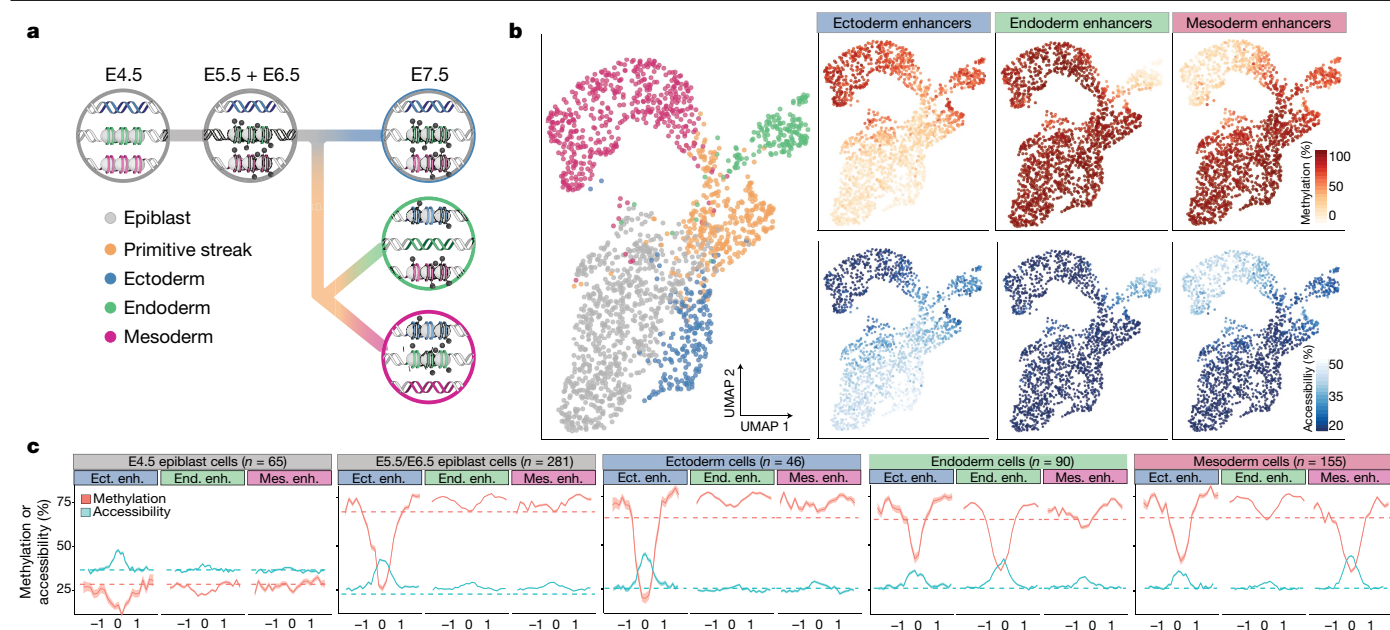


Fig. 3 | DNA methylation and chromatin accessibility dynamics at lineage-defining enhancers across development. **a**, Illustration of the hierarchical model of enhancer epigenetic dynamics associated with germ-layer commitment. **b**, UMAP projection based on the MOFA factors inferred using all embryonic cells ($n=1,928$). Main plot, cells are coloured by lineage. Smaller plots, cells are coloured by average methylation (top) or accessibility (bottom) at lineage-defining enhancers. For cells with RNA-expression data only, the MOFA factors were used to estimate the methylation and accessibility levels.

c, Profiles of methylation (red) and accessibility (blue) at lineage-defining enhancers (enh.) ($n=3,918$ for ectoderm, $n=1,930$ for endoderm, $n=1,417$ for mesoderm) across development. Running averages in 50-bp windows around the centre of the ChIP-seq peaks (2 kb upstream and downstream) are shown. Solid lines show the mean across cells and shaded areas represent the s.d. E5.5 and E6.5 epiblast cells show similar profiles and are combined. Dashed horizontal lines represent genome-wide background levels for methylation (red) and accessibility (blue).

endoderm enhancers. In both cases, changes in methylation and accessibility co-occur, suggesting tight co-regulation of the two epigenetic layers.

TET enzymes drive enhancer demethylation

TET methylcytosine dioxygenase enzymes have been implicated in enhancer demethylation^{23,24}, and loss-of-function experiments suggest that TET enzymes are vital for gastrulation^{25,26}. To test whether TET enzymes drive lineage-specific demethylation, we differentiated both wild-type ES cells and ES cells deficient for all three TET enzymes (*Tet* TKO) into embryoid bodies and analysed the cells using scNMT-seq.

Mapping the RNA-expression profiles to the *in vivo* gastrulation atlas shows that wild-type embryoid bodies recapitulate the transition from a pluripotent epiblast at day 2 of differentiation to the primitive streak between days 4 and 5 (Fig. 4a, b). At days 6 and 7, we observe the emergence of mature mesoderm structures including haematopoietic cell types (Fig. 4a, b, Extended Data Fig. 12). Expression of marker genes is restricted to the expected lineage and differential expression between lineages agrees with the *in vivo* results (Extended Data Fig. 12). Moreover, the global dynamics of DNA methylation and chromatin accessibility in wild-type embryoid bodies substantially mirror the *in vivo* data (Extended Data Fig. 12).

Comparison of wild type with *Tet* TKO differentiation in the epiblast-like cells at day 2 revealed higher DNA methylation in ectoderm enhancers in the *Tet* TKO cells, but no differences in mesoderm or endoderm enhancers (Fig. 4c). Re-analysis of methylation measurements from *Tet* TKO embryos confirms that the same pattern is observed *in vivo*²⁵ (Extended Data Fig. 12). Impaired demethylation is also associated with differences in differentiation timing, with *Tet* TKO cells showing an increased proportion of early mesendoderm differentiation at day 4 to 5 (Fig. 4a, b). However, at day 6 to 7 *Tet* TKO cells do not properly

demethylate lineage-specific enhancers and do not differentiate into mature mesodermal cell types (Fig. 4c).

These observations indicate that demethylation of lineage-defining enhancers is at least partially driven by TET proteins. Although enhancer demethylation does not seem to be required for early mesoderm commitment, the lack of haematopoietic cells in the *Tet* TKO cells suggests that demethylation may be important for subsequent lineage progression. Consistently, *Tet* TKO embryos are able to initiate gastrulation, but by E8.5 they display defects in mesoderm-derived cell types, including heart or somites²⁵.

Discussion

Our results show that pluripotent epiblast cells are epigenetically primed for an ectoderm fate as early as E4.5. This finding supports the existence of a 'default' path in Waddington's epigenetic landscape model, providing a potential mechanism for the phenomenon of 'default' differentiation of neuroectodermal tissue from ES cells^{27,28}. By contrast, endoderm and mesoderm are actively diverted from the default path by demethylation and chromatin opening at the corresponding enhancer elements^{17,24,25}. Thus, the germ-layer epigenome is defined during gastrulation by a hierarchical, or asymmetric, epigenetic model (Fig. 3a).

More generally, these results have important implications for the role of the epigenome in defining lineage commitment. We speculate that asymmetric epigenetic priming—whereby early progenitors are epigenetically primed for a default cell type—may be a more general feature of lineage commitment *in vivo*. In support of this hypothesis, two recent studies have identified default pathways in foregut specification and osteogenesis^{29,30}. Future studies that use multi-omics approaches to investigate cell populations have the potential to transform our understanding of cell-fate decisions, with important implications for stem cell biology.

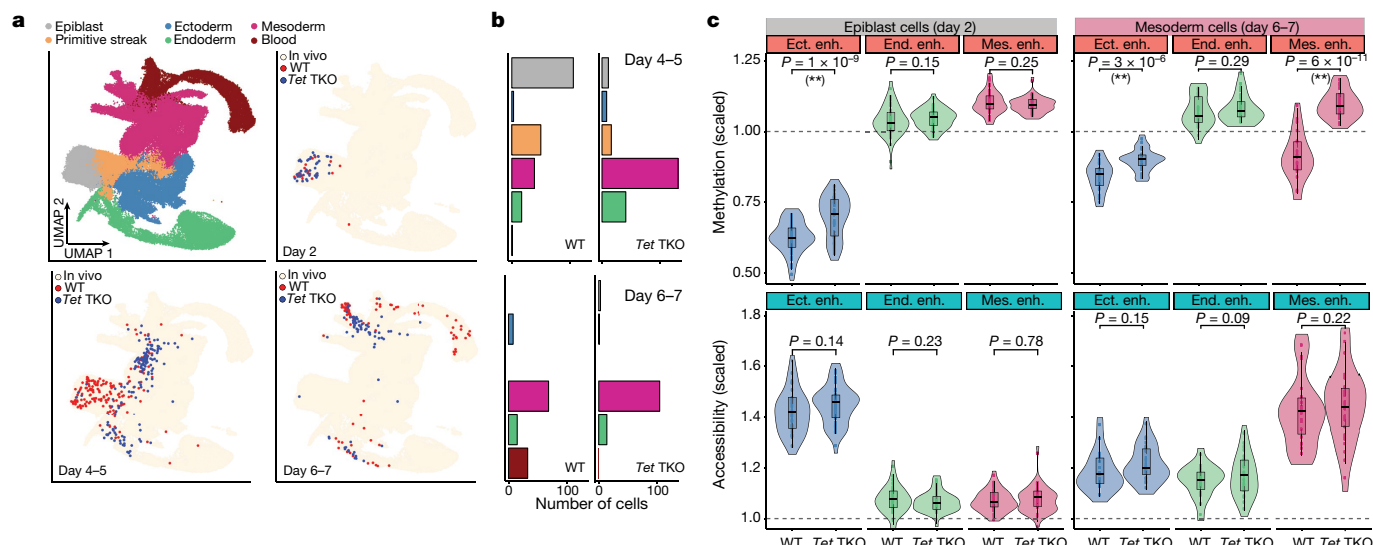


Fig. 4 | TET enzymes are required for efficient demethylation of mesoderm-defining enhancers and subsequent blood differentiation in embryoid bodies. **a**, UMAP projection of stages E6.5 to E8.5 of the atlas dataset (no extra-embryonic cells). Top left, cells coloured by lineage assignment. The remaining plots show, for different days of embryoid body differentiation, the nearest neighbours that were used to assign cell-type labels to the embryoid body dataset. Wild-type (WT) cells are red ($n = 438$), Tet TKO cells are blue ($n = 436$). We grouped days 4–5 and 6–7 together because of the similarity in the cell types recovered. **b**, Bar plots showing the numbers of each cell type for each

day of embryoid body differentiation, grouped by genotype ($n = 438$ WT and 436 KO). **c**, Overlaid box and violin plots show the distribution of DNA methylation (top) or chromatin accessibility (bottom) for lineage-defining enhancers in epiblast-like cells at day 2 ($n = 46$ (WT) and $n = 44$ (Tet TKO)) and mesoderm-like cells at days 6–7 ($n = 22$ (WT) and $n = 32$ (Tet TKO)). The yaxes show methylation or accessibility scaled to the genome-wide levels. Box plots show median levels and the first and third quartile, whiskers show 1.5x the interquartile range. P-values shown result from comparisons of group means (t-test). Asterisks denote significant differences (FDR < 10%).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1825-8>.

- Peng, G. et al. Spatial transcriptome for the molecular annotation of lineage fates and cell identity in mid-gastrula mouse embryo. *Dev. Cell* **36**, 681–697 (2016).
- Mohammed, H. et al. Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. *Cell Rep.* **20**, 1215–1228 (2017).
- Wen, J. et al. Single-cell analysis reveals lineage segregation in early post-implantation mouse embryos. *J. Biol. Chem.* **292**, 9840–9854 (2017).
- Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
- Chan, M. M. et al. Molecular recording of mammalian embryogenesis. *Nature* **570**, 77–82 (2019).
- Auclair, G., Guibert, S., Bender, A. & Weber, M. Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse. *Genome Biol.* **15**, 545 (2014).
- Lee, H. J., Hore, T. A. & Reik, W. Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell* **14**, 710–719 (2014).
- Zhang, Y. et al. Dynamic epigenomic landscapes during early lineage specification in mouse embryos. *Nat. Genet.* **50**, 96–105 (2018).
- Macaulay, I. C. et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
- Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nat. Biotechnol.* **33**, 285–289 (2015).
- Angermueller, C. et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13**, 229–232 (2016).
- Clark, S. J. et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).
- Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
- Smith, Z. D. et al. A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**, 339–344 (2012).
- Argelaguet, R. et al. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
- Xiang, Y. et al. Epigenomic analysis of gastrulation reveals a unique chromatin state for primed pluripotency. *Nat. Genet.* <https://doi.org/10.1038/s41588-019-0545-1> (2019).
- Cusanovich, D. A. et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **555**, 538–542 (2018).
- Daugherty, A. C. et al. Chromatin accessibility dynamics reveal novel functional enhancers in *C. elegans*. *Genome Res.* **27**, 2096–2107 (2017).
- Bogdanović, O. et al. Active DNA demethylation at enhancers during the vertebrate phylotypic period. *Nat. Genet.* **48**, 417–426 (2016).
- Kazakevych, J., Sayols, S., Messner, B., Krienke, C. & Soshnikova, N. Dynamic changes in chromatin states during specification and differentiation of adult intestinal stem cells. *Nucleic Acids Res.* **45**, 5770–5784 (2017).
- Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
- Kim, H. S. et al. Pluripotency factors functionally premark cell-type-restricted enhancers in ES cells. *Nature* **556**, 510–514 (2018).
- Rasmussen, K. D. & Helin, K. Role of TET enzymes in DNA methylation, development, and cancer. *Genes Dev.* **30**, 733–750 (2016).
- Sardina, J. L. et al. Transcription factors drive Tet2-mediated enhancer demethylation to reprogram cell fate. *Cell Stem Cell* **23**, 727–741.e9 (2018).
- Dai, H.-Q. et al. TET-mediated DNA demethylation controls gastrulation by regulating Lefty–Nodal signalling. *Nature* **538**, 528–532 (2016).
- Li, X. et al. Tet proteins influence the balance between neuroectodermal and mesodermal fate choice by inhibiting Wnt signaling. *Proc. Natl Acad. Sci. USA* **113**, E8267–E8276 (2016).
- Tropepe, V. et al. Direct neural fate specification from embryonic stem cells: a primitive mammalian neural stem cell stage acquired through a default mechanism. *Neuron* **30**, 65–78 (2001).
- Muñoz-Sanjuán, I. & Brivanlou, A. H. Neural induction, the default model and embryonic stem cells. *Nat. Rev. Neurosci.* **3**, 271–280 (2002).
- Rauch, A. et al. Osteogenesis depends on commissioning of a network of stem cell transcription factors that act as repressors of adipogenesis. *Nat. Genet.* **51**, 716–727 (2019).
- Banerjee, K. K. et al. Enhancer, transcriptional, and cell fate plasticity precedes intestinal determination during endoderm development. *Genes Dev.* **32**, 1430–1442 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Embryos and single cell isolation

All mice used in this study were C57BL/6Bab and were bred and maintained in the Babraham Institute Biological Support Unit. Ambient temperature was about 19–21 °C and relative humidity was 52%. Lighting was provided on a 12 h:12 h light:dark cycle, including 15 min ‘dawn’ and ‘dusk’ periods of subdued lighting. After weaning, mice were transferred to individually ventilated cages with 1–5 mice per cage. Mice were fed CRM (P) VP diet (Special Diet Services) ad libitum and received seeds (for example, sunflower or millet) at the time of cage-cleaning as part of their environmental enrichment. All mouse experimentation was approved by the Babraham Institute Animal Welfare and Ethical Review Body. Animal husbandry and experimentation complied with existing European Union and United Kingdom Home Office legislation and local standards. Sample sizes were determined to obtain at least 50 cells for each germ layer. No randomization or blinding was performed. Sex of embryos was not known at the time of collection. Single-cells from E4.5 to E5.5 embryos were collected as previously described². E6.5 and E7.5 embryos were dissected to remove extra-embryonic tissues and dissociated in TrypLE for 10 min at room temperature. Undigested portions were physically removed and the remainder filtered through a 30-µm filter before isolation using flow cytometry.

Tet TKO cell culture

Tet1^{−/−} *Tet2*^{−/−} *Tet3*^{−/−} (C57BL6/129/FVB) and matching wild-type mouse ES cells³¹ were cultured in 2i+LIF medium (50/50 DMEM-F12 (Gibco, 31330-038) and Neurobasal medium (Gibco, 21103-49) with serum-free N2B27 (0.5% N2 and 1% B27; Gibco), 0.1 mM 2-mercaptoethanol (Life Technologies, 31350-010) and 2 mM L-glutamine (Life Technologies, 25030-024) supplemented with LIF, MEK inhibitor PD0325901 (1 µM) and GSK3 inhibitor CHIR99021 (3 µM), all from Department of Biochemistry, University of Cambridge). ES cells were cultured on tissue culture plastic pre-coated with 0.1% gelatine in H₂O and were passaged when approaching confluence (every 2–3 days).

For the embryoid body differentiation assay, 2 × 10⁴ ES cells were collected in medium consisting of DMEM (Life Technologies, 10566-016), 15% fetal bovine serum (Gibco, 10270106), 1 × non-essential amino acids (NEAA) (Life Technologies, 11140050), 0.1 mM 2-mercaptoethanol (Life Technologies, 31350-010), 2 mM L-glutamine (Life Technologies, 25030-024) in ultra-low attachment 96-well plates (Sigma-Aldrich, CLS7007). All cells were cultured in a humidified incubator at 37 °C in 5% CO₂ and 20% O₂. Embryoid bodies were collected 2, 4, 5, 6 and 7 days after induction of differentiation and dissociated into single cells using accutase before flow sorting. Cell lines were subject to routine mycoplasma testing using the MycoAlert testing kit (Lonza) and tested negative. Cell lines were not authenticated.

scNMT-seq library preparation

Single cells were flow-sorted (E6.5 and E7.5 stages, using a BD Influx or BD Aria III) or manually picked when cell numbers were too low (E4.5, E5.5). Cells were isolated into 96-well PCR plates containing 2.5 µl of methylase reaction buffer (1 × M.CviPI Reaction buffer (NEB), 2 U M.CviPI (NEB), 160 µM S-adenosylmethionine (NEB), 1 U µl^{−1} RNasein (Promega), 0.1% IGEPAL CA-630 (Sigma)). Samples were incubated for 15 min at 37 °C to methylate accessible chromatin before the reaction was stopped with the addition of RLT plus buffer (Qiagen) and samples frozen down and stored at −80 °C before processing. Poly-A RNA was captured on oligo-dT conjugated to magnetic beads and amplified cDNA was prepared according to the G&T-seq³² and Smartseq2 protocols³³. The lysate containing gDNA was purified on AMPureXP beads

before bisulfite-sequencing (BS-seq) libraries were prepared according to the scBS-seq protocol³⁴.

A subset of embryo cells were processed for scRNA-seq only (1,419 cells after QC). These followed the same protocol but we discarded the gDNA after separation.

A full step-by-step protocol for scNMT-seq is available at <https://doi.org/10.17504/protocols.io.6jnhcme>.

Sequencing

All sequencing was carried out on a NextSeq500 instrument. BS-seq libraries were sequenced in 48-plex pools using 75-bp paired-end reads in high-output mode. RNA-seq libraries were pooled as either 384 plexes and sequenced using 75-bp paired-end reads in high-output mode or 192 plexes and sequenced using 75-bp paired-end reads in mid-output mode. This yielded a mean raw sequencing depth of 8.5 million (BS-seq) and 1 million (RNA-seq) paired-end reads per cell.

RNA-seq alignment and quantification

RNA-seq libraries were aligned to the GRCm38 mouse genome build using HiSat2³⁵ (v.2.1.0) using options `-dta -sp. 1000,1000 -no-mixed -no-discordant`, yielding a mean of 681,000 aligned reads per cell. Subsequently, gene expression counts were quantified from the mapped reads using featureCounts³⁶ with the Ensembl gene annotation³⁷ (v.87). Only protein-coding genes matching canonical chromosomes were considered. The read counts were log-transformed and size-factor adjusted³⁸.

BS-seq alignment and methylation/accessibility quantification

BS-seq libraries were aligned to the bisulfite converted GRCm38 mouse genome using Bismark³⁹ (v.0.19.1) in single-end nondirectional mode. Following the removal of PCR duplicates, we retained a mean of 1.6 million reads per cell. Methylation calling and separation of endogenous methylation (from A-C-G and T-C-G trinucleotides) and chromatin accessibility (G-C-A, G-C-C and G-C-T trinucleotides) was performed with Bismark using the `-NOME` option of the coverage2cytosine script.

Following a previous approach⁴⁰, individual CpG or GpC sites in each cell were modelled using a binomial distribution in which the number of successes is the number of reads that support methylation and the number of trials is the total number of reads. A CpG methylation or GpC accessibility rate for each site and cell was calculated by maximum likelihood. The rates were subsequently rounded to the nearest integer (0 or 1).

When aggregating over genomic features, CpG methylation and GpC accessibility rates were computed assuming a binomial model, with the number of trials being the number of observed CpG sites and the number of successes being the number of methylated CpGs. Notably, this implies that DNA methylation and chromatin accessibility is quantified as a rate (or a percentage). We avoid binarizing DNA methylation and chromatin accessibility values into low and high states, as this is not a good representation of the continuous nature of the data (Extended Data Fig. 3).

ChIP-seq data processing

ChIP-seq data were obtained from the Gene Expression Omnibus accession code GSE125318. Reads were trimmed using Trim Galore (v.0.4.5, cutadapt 1.15, single end mode) and mapped to *Mus musculus* GRCm38 using Bowtie2⁴¹ (v.2.3.2). Read 2 was excluded from the analysis for paired-end samples because of low-quality scores (Phred <25). All analyses were performed using SeqMonk (<https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/>). For quantification, read length was extended to 300 bp and regions of coverage outliers and extreme strand bias were excluded as these were assumed to be alignment artefacts. Comparison of datasets with different read lengths did not reveal major mapping differences, and thus mapped, extended reads were merged for samples that were sequenced across more than one lane.

Samples were similar overall regarding total mapped read numbers, distribution of reads and ChIP enrichment.

To best represent the underlying ChIP-seq signal, different methods to define enriched genomic regions were used for H3K4me3 and H3K27ac marks. For H3K4me3, a SeqMonk implementation of MACS⁴² with the local rescoring step omitted was used ($P < 10^{-15}$, fragment size 300 bp), and enriched regions closer than 100 bp were merged. Peaks were called separately for each lineage. For H3K27ac, reads were quantitated per 500-bp tiles correcting per million total reads and excluding duplicate reads. Smoothing subtraction quantification was used to identify local maxima (value > 1), and peaks closer than 500 bp apart were merged. Lineage-specific peak annotations exclude peaks that are also present in one of the other lineages, and only peaks present in both replicates were considered (Extended Data Fig. 5).

Publicly available ChIP-seq libraries for H3K27ac^{20–22} were processed with Trim Galore and Bowtie2 (see above), and analysed in Seqmonk. Read counts were determined for 1-kb non-overlapping tiles and, separately, for lineage-specific enhancers (average length 1.2 kb). The genomic tiles were used to determine the distribution of H3K27ac across the genome. Enhancers were classified as marked if their read counts were within the top 5% of the distribution.

scRNA-seq and scBS-seq quality control

For RNA expression, cells with less than 100,000 mapped reads and with less than 500 expressed genes were excluded. For DNA methylation and chromatin accessibility, cells with less than 50,000 CpG sites and 500,000 GpC sites covered, respectively, were discarded (Extended Data Fig. 1).

Lineage assignment using RNA expression

Lineages were assigned by mapping the RNA-expression profiles to a comprehensive single-cell atlas from the same stages⁴, when available (stages E6.5 and E7.5), or by SC3⁴³ otherwise (stages E4.5 and E5.5) (Extended Data Fig. 2). Extra-embryonic cells were identified by these methods and excluded from further analyses.

The mapping was performed by matching mutual nearest neighbours⁴⁴. First, count matrices from both experiments were concatenated and normalized together. Highly variable genes were selected³⁸ from the resulting expression matrix and were used as input for principal components analysis. Subsequently, batch correction was applied to remove the technical variability between the two experiments and a k -nearest neighbours graph was computed between them. For each scNMT-seq cell, the cell type was selected as the mode from a Dirichlet distribution given by the cell type distribution of the top 30 nearest neighbours in the atlas (that is, majority voting).

Correlation analysis

To identify genes with an association between the mRNA expression and promoter epigenetic status, we calculated the correlation coefficient for each gene across all cells between the RNA expression and the corresponding DNA methylation or chromatin accessibility levels at the gene's promoter ± 2 kb around the transcription start site (TSS).

As a filtering criterion, we required, for each genomic feature, a minimum number of 1 CpG (methylation) or 5 GpC (accessibility) measurements in at least 50 cells. Additionally, the top-5,000 most variable genes (across all cells) were selected, according to the rationale of independent filtering⁴⁵. Two-tailed Student's t -tests were performed to test for evidence against the null hypothesis of no correlation, and P values were adjusted for multiple testing using the Benjamini–Hochberg procedure⁴⁶.

Differential DNA methylation and chromatin accessibility analysis

Differential analysis of DNA methylation and chromatin accessibility was performed using a Fisher exact test independently for each

genomic element. Cells were aggregated into two exclusive groups and, for a given genomic element, we created a contingency table by aggregating (across cells) the number of methylated and unmethylated nucleotides. Multiple testing correction was applied using the Benjamini–Hochberg procedure. As a filtering criteria, we required 1 CpG (methylation) and 5 GpC (accessibility) observations in at least 10 cells per group. Non-variable regions were filtered out before differential testing.

Motif enrichment

To find transcription factor motifs enriched in lineage-associated sites, we used H3K27ac sites that were identified as differentially accessible between lineages as explained above. We tested for enrichment over a background of all H3K27ac sites using *ame* (meme suite⁴⁷ v.4.10.1) with parameters –method fisher –scoring avg. Position frequency matrices were downloaded from the Jasp core vertebrates database⁴⁸. This is a curated list of experimentally derived binding motifs and not an exhaustive set, which means that some important transcription factors will not be analysed, owing to absence of their motifs.

Differential RNA-expression analysis

Differential RNA-expression analysis between prespecified groups of interest was performed using the genewise negative binomial generalized linear model with quasi-likelihood test from edgeR⁴⁹. Significant hits were called with a 1% FDR (Benjamini–Hochberg procedure) and a minimum \log_2 fold change of 1. Genes with low expression (mean \log_2 counts < 0.5) were filtered out before differential testing⁴⁵.

Dimensionality reduction for DNA methylation and chromatin accessibility data using Bayesian factor analysis

To handle the large number of missing values in DNA methylation and chromatin accessibility data, we used a linear Bayesian factor analysis model¹⁵. The linearity assumption renders the model output directly interpretable, and more robust to changes in hyperparameters than nonlinear methods, particularly with small numbers of cells. We trained every model using the top-5,000 most variable features and we constrained the latent space to two latent factors, which were used for visualization (Fig. 1c, d, Extended Data Fig. 3). Variance-explained estimates were computed using the coefficient of determination as previously described¹⁵.

MOFA

The input to MOFA is a list of matrices, in which each matrix represents a different data modality. RNA-expression measurements were defined as one data modality. For DNA methylation and chromatin accessibility, we defined separate matrices for promoters, distal H3K27ac sites (enhancers) and H3K4me3 (TSS). Promoters were defined as a bidirectional 2-kb window around the TSS of protein-coding genes. For each genomic context, we created a DNA methylation matrix and a chromatin accessibility matrix by quantifying M -values for each cell and genomic element.

As a filtering criterion, genomic features were required to have a minimum of 1 CpG (methylation) or 5 GpC (accessibility) observed in at least 25 cells. Genes were required to have a minimum cellular detection rate of 25%. In addition, to reduce computational complexity, the top 1,000 most variable features were selected per view. Similarly, the top 2,500 most variable genes were selected for RNA expression.

Similar to most latent dimensionality reduction methods, the optimization procedure of MOFA is not guaranteed to find a global optimum. Following ref.¹⁵, model selection was performed by selecting the model with the highest evidence lower bound out of ten trials.

The number of factors was calculated by requiring a minimum of 1% variance explained in the RNA. The robustness of factors across trials was assessed by calculating the correlation coefficients between every

Article

pair of factors across the ten trials. All inferred factors were consistently found in all model instances.

The downstream characterization of the model output included several analyses. (1) Variance decomposition: quantification of the fraction of variance explained (R^2) by each factor in each view, using a coefficient of determination¹⁵. (2) Visualization of weights/loadings: the model learns a weight for every feature in each factor, which can be interpreted as a measure of feature importance. Features with large weights (in absolute value) are highly correlated with the factor values. (3) Visualization of factors: each MOFA factor captures a different dimension of cellular heterogeneity. All together, they define a latent space that maximizes the variance explained in the data (under some important sparsity assumptions¹⁵). The cells can be visualized in the latent space by plotting scatter plots of combinations of factors. (4) Gene set enrichment analysis: when inspecting the weights for a given factor, multiple features can be combined into a gene set-based annotation. For a given gene set G , we evaluate its significance via a parametric t -test (two-sided), whereby we compare the mean of the weights of the foreground set (features that belong to the set G) with the mean of the weights in the background set (features that do not belong to the set G). Resulting P values are adjusted for multiple testing using the Benjamini–Hochberg procedure from which significant pathways are called (FDR <10%).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Raw sequencing data together with processed files (RNA counts, CpG methylation reports, CpC accessibility reports) are available in the Gene Expression Omnibus under accession number GSE121708. Processed data can be downloaded from ftp://ftp.ebi.ac.uk/pub/databases/scnmt_gastrulation.

Code availability

All code used for analysis is available at https://github.com/rargelaguet/scnmt_gastrulation.

31. Hu, X. et al. Tet and TDG mediate DNA demethylation essential for mesenchymal-to-epithelial transition in somatic cell reprogramming. *Cell Stem Cell* **14**, 512–522 (2014).
32. Macaulay, I. C. et al. Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. *Nat. Protoc.* **11**, 2081–2103 (2016).
33. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
34. Clark, S. J. et al. Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nat. Protoc.* **12**, 534–547 (2017).
35. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
36. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
37. Yates, A. et al. Ensembl 2016. *Nucleic Acids Res.* **44** (D1), D710–D716 (2016).
38. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000 Res.* **5**, 2122 (2016).
39. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
40. Smallwood, S. A. et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).

41. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
42. Zhang, Y. et al. Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, R137 (2008).
43. Kiselev, V. Y. et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483–486 (2017).
44. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
45. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl Acad. Sci. USA* **107**, 9546–9551 (2010).
46. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B* **57**, 289–300 (1995).
47. McLeay, R. C. & Bailey, T. L. Motif enrichment analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* **11**, 165 (2010).
48. Khan, A. et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46** (D1), D260–D266 (2018).
49. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
50. Ohnishi, Y. et al. Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. *Nat. Cell Biol.* **16**, 27–37 (2014).
51. Yeom, Y. I. et al. Germline regulatory element of Oct-4 specific for the totipotent cycle of embryonal cells. *Development* **122**, 881–894 (1996).
52. Kalantry, S. et al. The amnionless gene, essential for mouse gastrulation, encodes a visceral-endoderm-specific protein with an extracellular cysteine-rich domain. *Nat. Genet.* **27**, 412–416 (2001).
53. Creighton, M. P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107**, 21931–21936 (2010).
54. Liang, G. et al. Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. *Proc. Natl Acad. Sci. USA* **101**, 7357–7362 (2004).
55. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
56. Scialdone, A. et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54–61 (2015).

Acknowledgements R.A. is a member of Robinson College at the University of Cambridge. We thank K. Tabbada, C. Murnane and N. Forrester of the Babraham Next Generation Sequencing Facility for assistance with Illumina sequencing; members of the Babraham Flow Cytometry Core Facility for cell sorting and the Babraham Biological Support Unit for animal work; Y. Zhang for help in processing the ChIP-seq data. L.C.S. was supported by an EMBO postdoctoral fellowship (ALTF 417-2018) and is currently a Marie Skłodowska-Curie fellow funded by the European Commission under the H2020 Programme. J.C.M. is supported by core funding from EMBL and CRUK. R.A. is supported by the EMBL International Predoc Programme. X.I.-S. is supported by Wellcome Trust Grant 108438/E/15/Z. F.B. is supported by the UK Medical Research Council (Career Development Award MR/M01536X/1). B.G. and J.N. are supported by core funding by the MRC and Wellcome Trust to the Wellcome–MRC Cambridge Stem Cell Institute. W.R. is supported by Wellcome (105031/Z/14/Z; 210754/Z/18/Z) and BBSRC (BBS/E/B/000C0422). O.S. is supported by core funding from EMBL and DKFZ and the EU (ERC project DECODE 810296).

Author contributions H.M., W.D. and W.R. conceived the project. S.S. and H.M. designed the study and generated pilot data. W.D., J.N. and L.C.S. performed embryo dissections and single-cell isolation. L.C.S. and T.L. performed in vitro differentiation experiments. S.J.C. and H.M. performed scNMT-seq library preparation. F.K. processed and managed sequencing data. C.K. analysed ChIP-seq datasets with assistance from Y.X. and C.W.H. R.A. and S.J.C. performed pre-processing and quality control of scNMT-seq data. R.A. and I.I.-R. mapped cells to the scRNA-seq atlas. R.A., S.J.C., F.B., L.C.S., X.I.-S., C.-A.K. and C.K. performed computational analysis. R.A. generated figures. R.A., S.J.C., L.C.S., O.S., J.C.M. and W.R. interpreted results and drafted the manuscript. G.S., P.J.R.-G., W.X., G.K., O.S., B.G., J.C.M. and W.R. supervised the project. All authors read and approved the final manuscript.

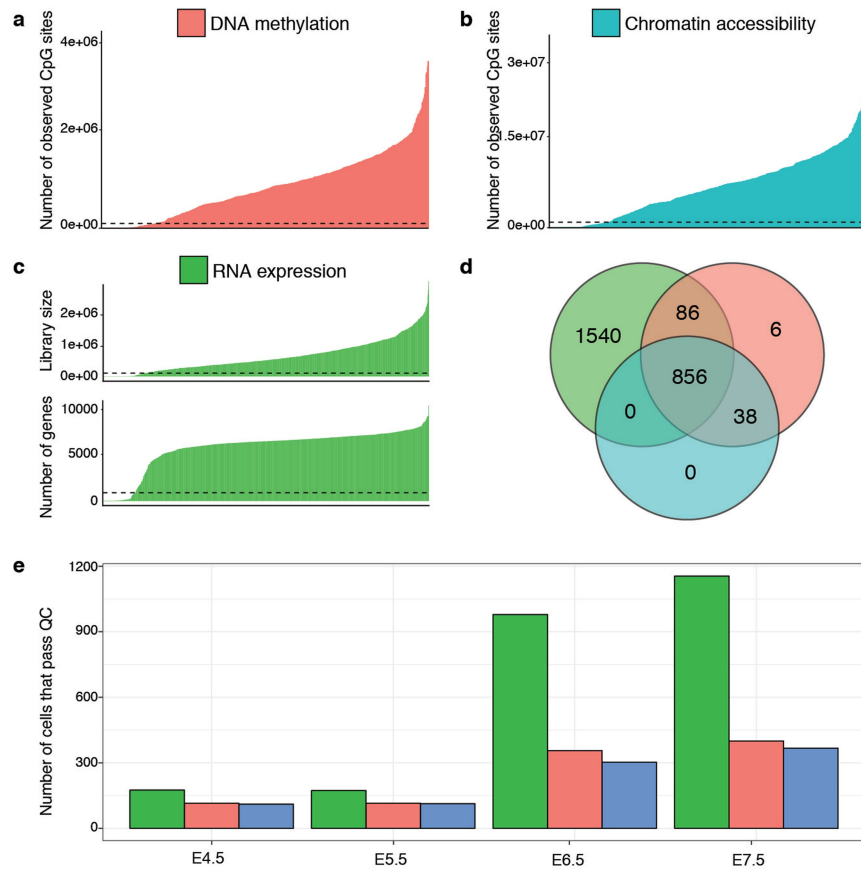
Competing interests W.R. is a consultant and shareholder of Cambridge Epigenetix. The remaining authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1825-8>.

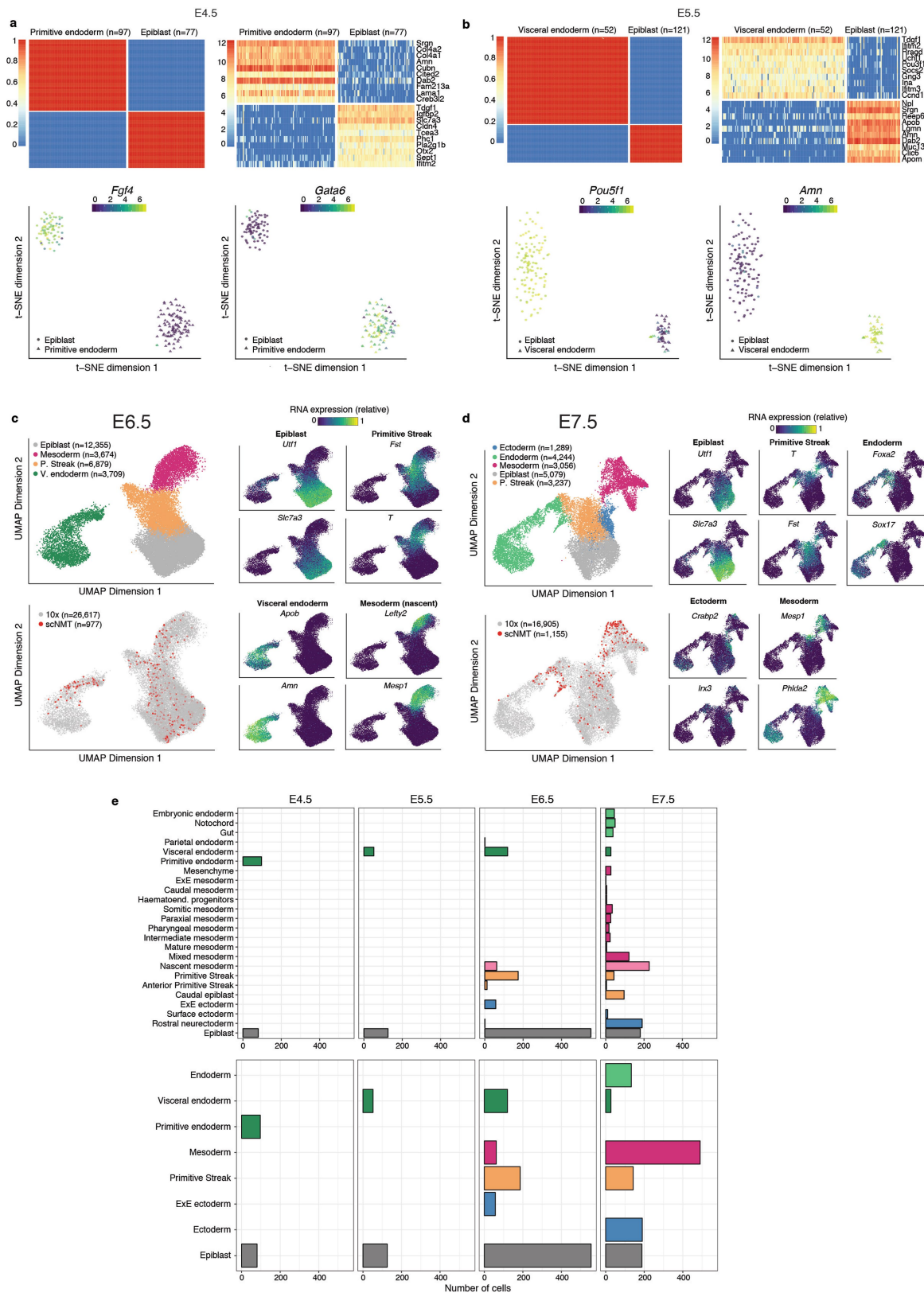
Correspondence and requests for materials should be addressed to S.J.C., O.S., J.C.M. or W.R. **Peer review information** Nature thanks Andrew Adey and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | scNMT-seq quality controls. **a, b**, Number of observed cytosines in CpG (red; **a**) or GpC (blue; **b**) contexts respectively. Each bar corresponds to one cell. Cells are sorted by total number of CpG or GpC sites. Cells below the dashed line were discarded on the basis of poor coverage ($n=1,105$). **c**, RNA-library size per cell. Top, total number of reads. Bottom, number of expressed genes (read counts >0). Cells below the dashed line were

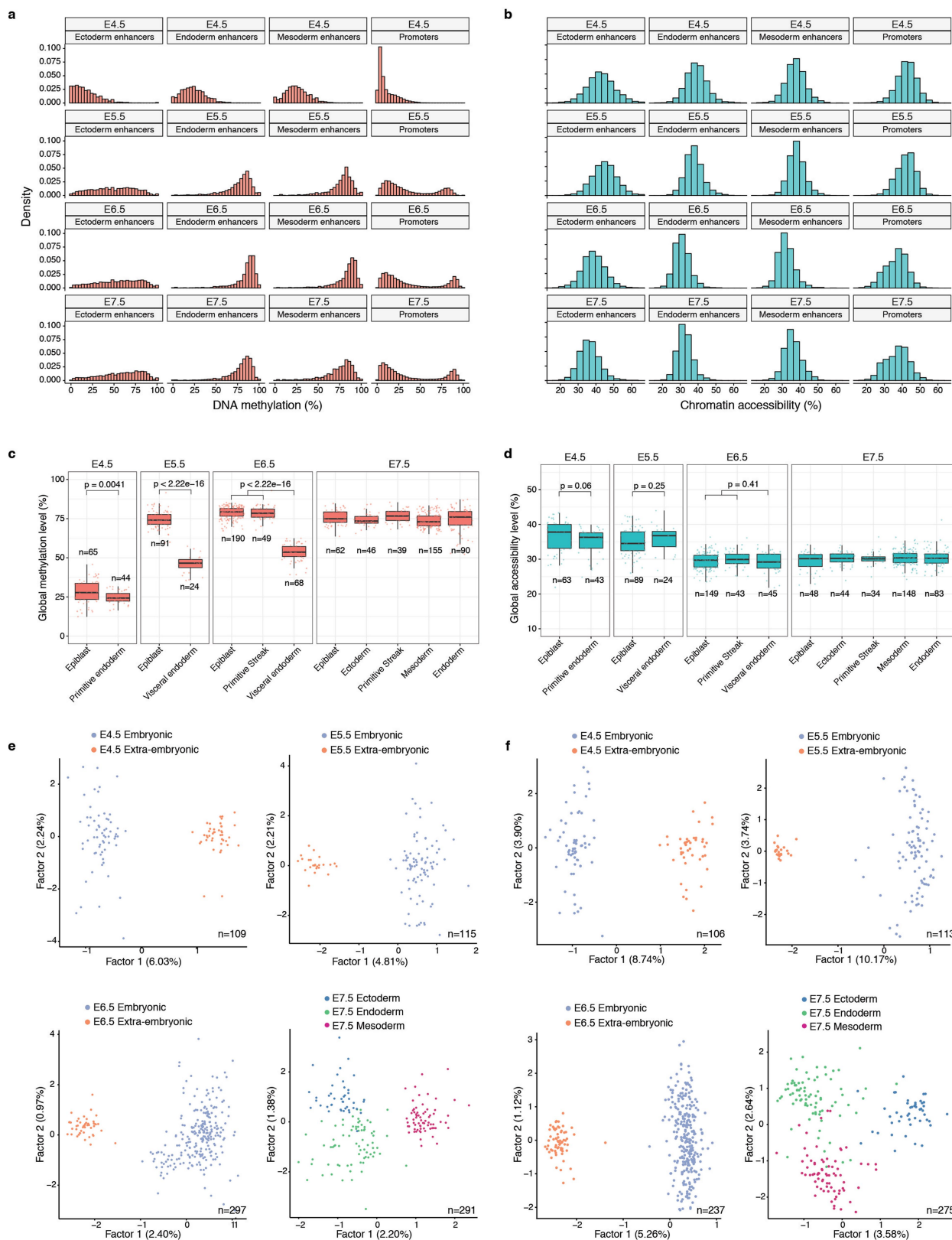
discarded on the basis of poor coverage ($n=2,524$). **d**, Venn diagram displaying the number of cells that pass quality control for RNA expression (green), DNA methylation (red) and chromatin accessibility (blue). **e**, Number of cells that pass quality control for each molecular layer, grouped by stage. For 1,419 out of 2,524 total cells, only the RNA expression was sequenced.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Cell-type assignments based on RNA expression. a, b, Lineage assignment of E4.5 cells (**a**; $n = 175$) and E5.5 cells (**b**; $n = 173$). Top left, SC3 consensus plots representing the similarity between cells on the basis of averaging of clustering results from multiple combinations of clustering parameters. Top right, heat map showing the RNA expression (log normalized counts) of the ten most informative gene markers for each cluster. Bottom left, t -distributed stochastic neighbour embedding (t -SNE) representation of the RNA-expression data coloured by the expression of *Fgf4* and *Pou5f1*, known E4.5 and E5.5 epiblast markers^{50,51}, respectively. Bottom right, t -SNE representation of the RNA-expression data coloured by the expression of *Gata6* and *Amn*, known E4.5 primitive endoderm and E5.5 visceral endoderm

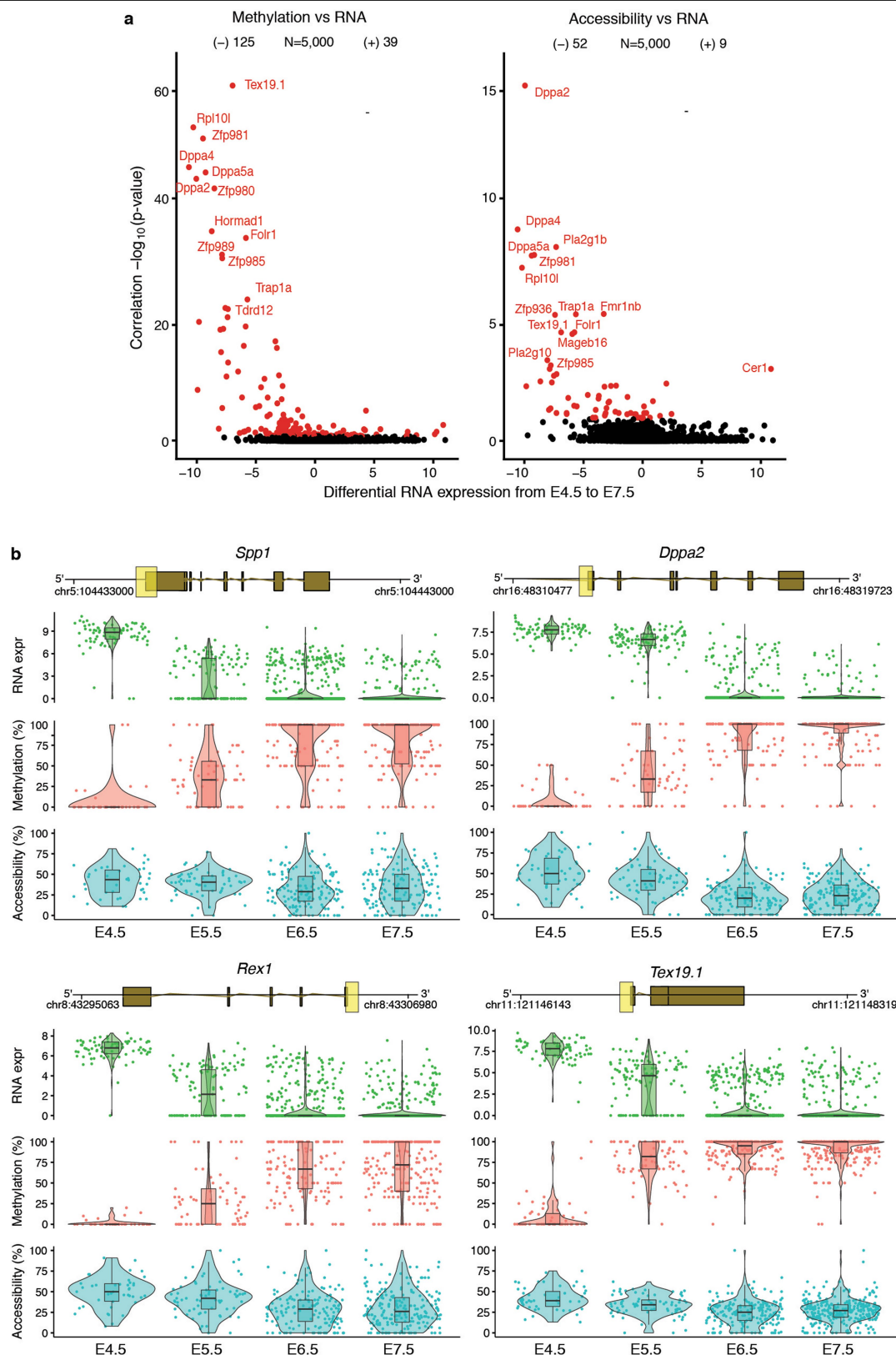
markers⁵². **c, d,** Lineage assignment of E6.5 cells (**c**; $n = 977$) and E7.5 cells (**d**; $n = 1,155$). Left, UMAP projection of the atlas dataset (stages E6.5 to E7.0 to assign E6.5 cells and E7.0 to E8.0 to assign E7.5 cells). In the top-left panel, cells are coloured by lineage assignment. In the bottom-left panel, the cells coloured in red are the nearest neighbours that were used to transfer labels to the scNMT-seq dataset. In right panels, cells are coloured by the relative RNA expression of lineage-marker genes. **e,** Top, number of cells per lineage, using the maximally resolved cell types reported in ref. ⁴. Bottom, number of cells per lineage after aggregation of cell types belonging to the same germ layer or extra-embryonic tissue type, as used in this study.



Extended Data Fig. 3 | See next page for caption.

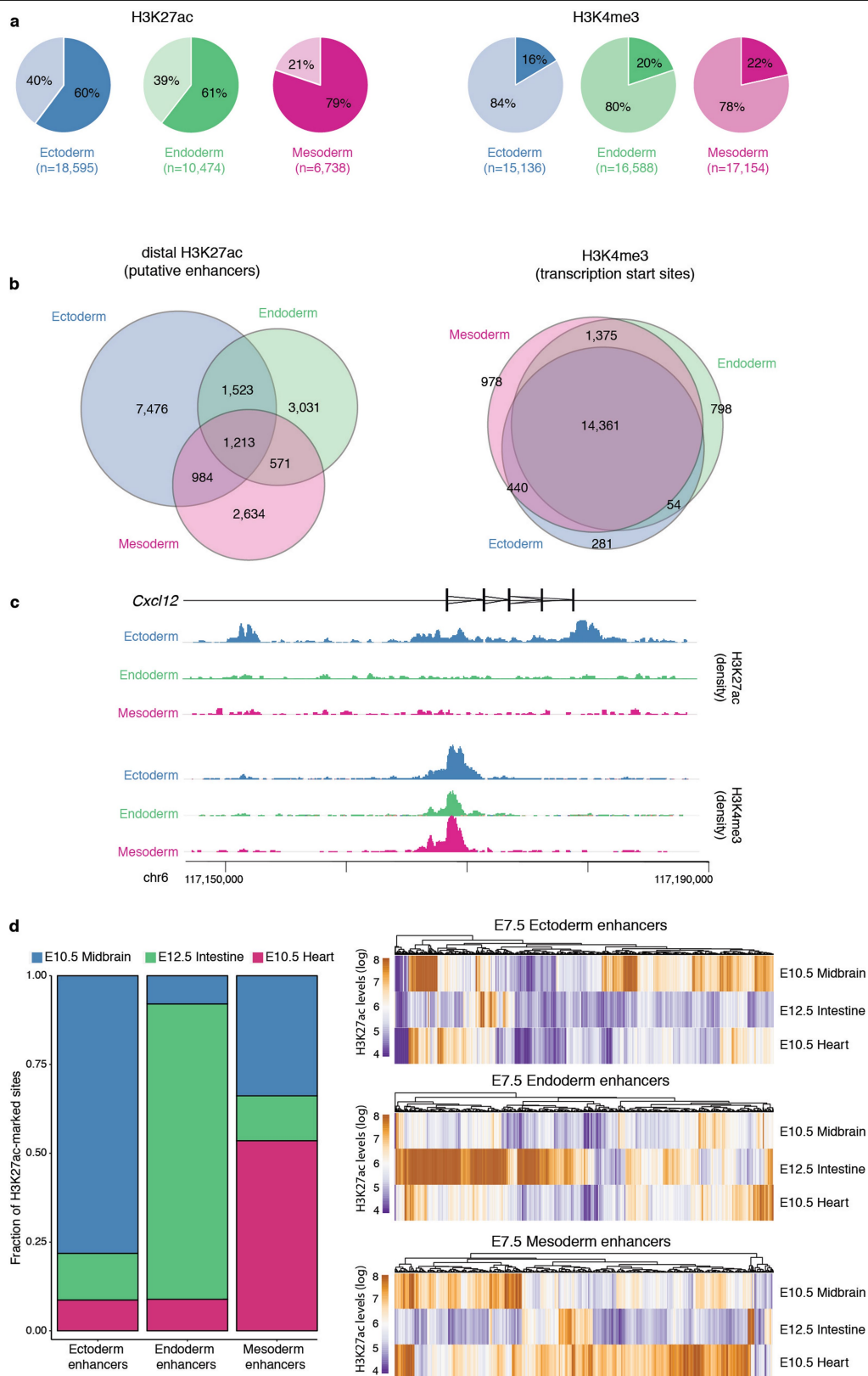
Extended Data Fig. 3 | Global methylation and chromatin accessibility dynamics. a, b, Distribution of DNA methylation (**a**) and chromatin accessibility levels (**b**) per stage and genomic context. When aggregating over genomic features, CpG methylation and GpC accessibility levels (%) are computed assuming a binomial model, with the number of trials being the total number of observed CpG (or GpC) sites and the number of successes being the number of methylated CpG (or GpC) sites (Methods). Notably, this implies that DNA methylation and chromatin accessibility are quantified as a percentage and are not binarized into low or high states. As this figure shows, the distribution of DNA methylation and chromatin accessibility across loci (after aggregating measurements across all cells per stage) is largely continuous and does not show bimodality. Hence, a binary approach similar to that sometimes used for differentiated cell types would not provide a good representation of the data. **c, d,** Box plots showing the distribution of genome-wide CpG methylation levels (**c**) or GpC accessibility levels (**d**) per stage and lineage. Each dot represents a single cell. Box plots show median levels and the first and third

quartile, whiskers show $1.5 \times$ the interquartile range. At a significance threshold of 0.01 (*t*-test, two-sided), the global DNA methylation levels differ between embryonic and extra-embryonic lineages, but the global chromatin accessibility levels do not. **e, f,** Dimensionality reduction of DNA methylation (**e**) and chromatin accessibility (**f**) data. To perform dimensionality reduction while handling the large amount of missing values, we used a Bayesian factor analysis model (Methods). Scatter plots of the first two latent factors (sorted by variance explained) for models trained with cells from the indicated stages are shown. From E4.5 to E6.5, cells are coloured by embryonic and extra-embryonic origin. At E7.5, cells are coloured by the primary germ layer. All lineage assignments were made using the cells' corresponding RNA-expression levels (Extended Data Fig. 2). The fraction of variance explained by each factor is displayed in parentheses. The input data were *M*-values quantified over DNase I hypersensitive sites profiled in ES cells ($n = 175,231$, subset to the top 5,000 most variable sites to fit the model).



Extended Data Fig. 4 | DNA methylation and chromatin accessibility changes in promoters are associated with repression of early pluripotency and germ cell markers. **a**, Volcano plots display differential RNA-expression levels between E4.5 and E7.5 cells (in \log_2 counts, x-axis) versus adjusted correlation P values (FDR $< 10\%$ in red, Benjamini-Hochberg correction, $n = 5,000$ genes). Left, DNA methylation versus RNA-expression correlations; right, chromatin accessibility versus RNA expression. Negative values for differential RNA expression indicate higher expression in E4.5, whereas

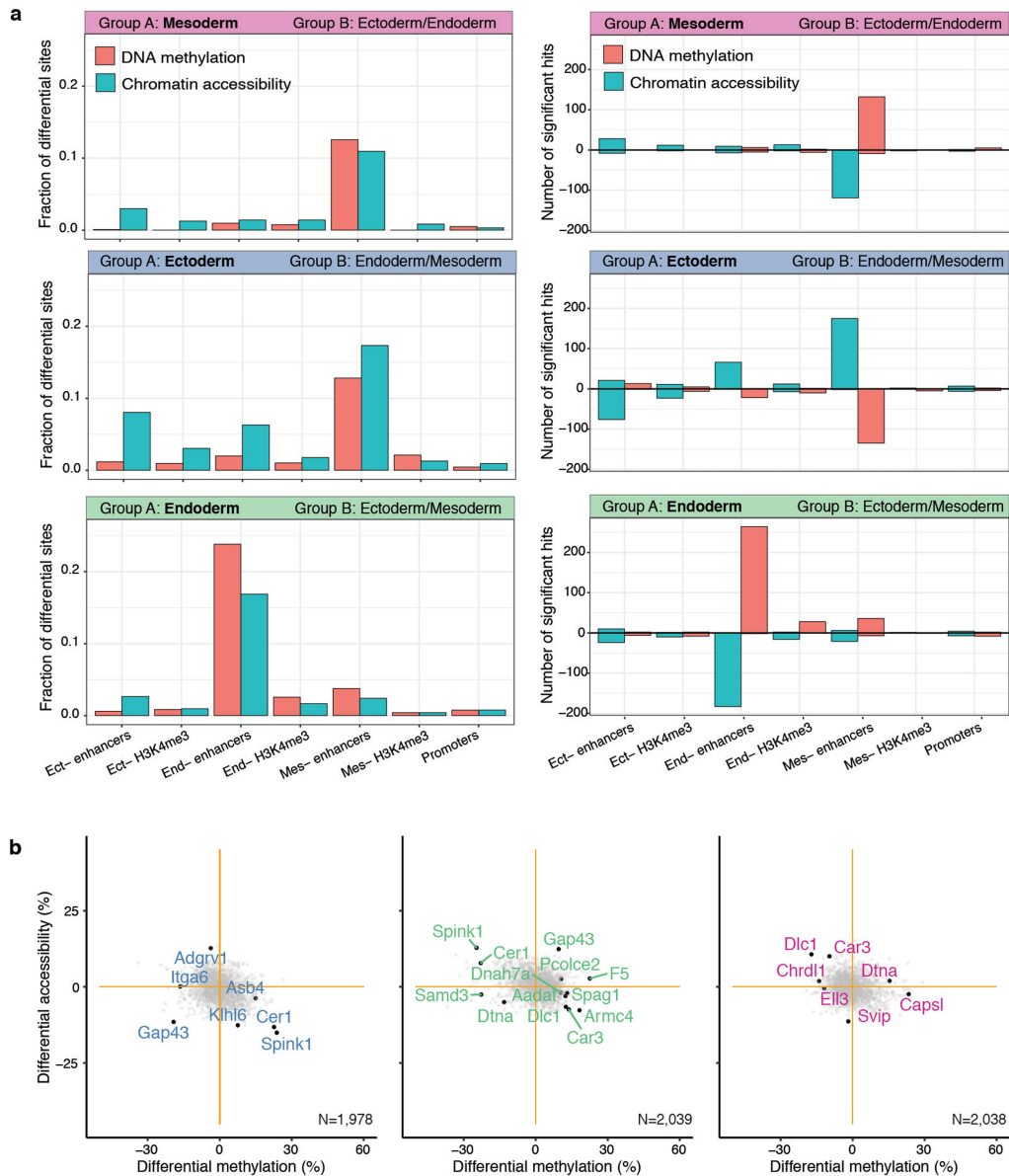
positive values indicate higher expression in E7.5. **b**, Illustrative examples of epigenetic repression of early pluripotency and germ cell markers. Box and violin plots show the distribution of RNA expression (\log_2 counts, green), DNA methylation (red) and chromatin accessibility (blue) levels per stage. Box plots show median coverage and the first and third quartile, whiskers show $1.5 \times$ the interquartile range. Each dot corresponds to one cell. For each gene a genomic track is shown on top, and the promoter region that is used to quantify DNA methylation and chromatin accessibility levels is highlighted in yellow.



Extended Data Fig. 5 | See next page for caption.

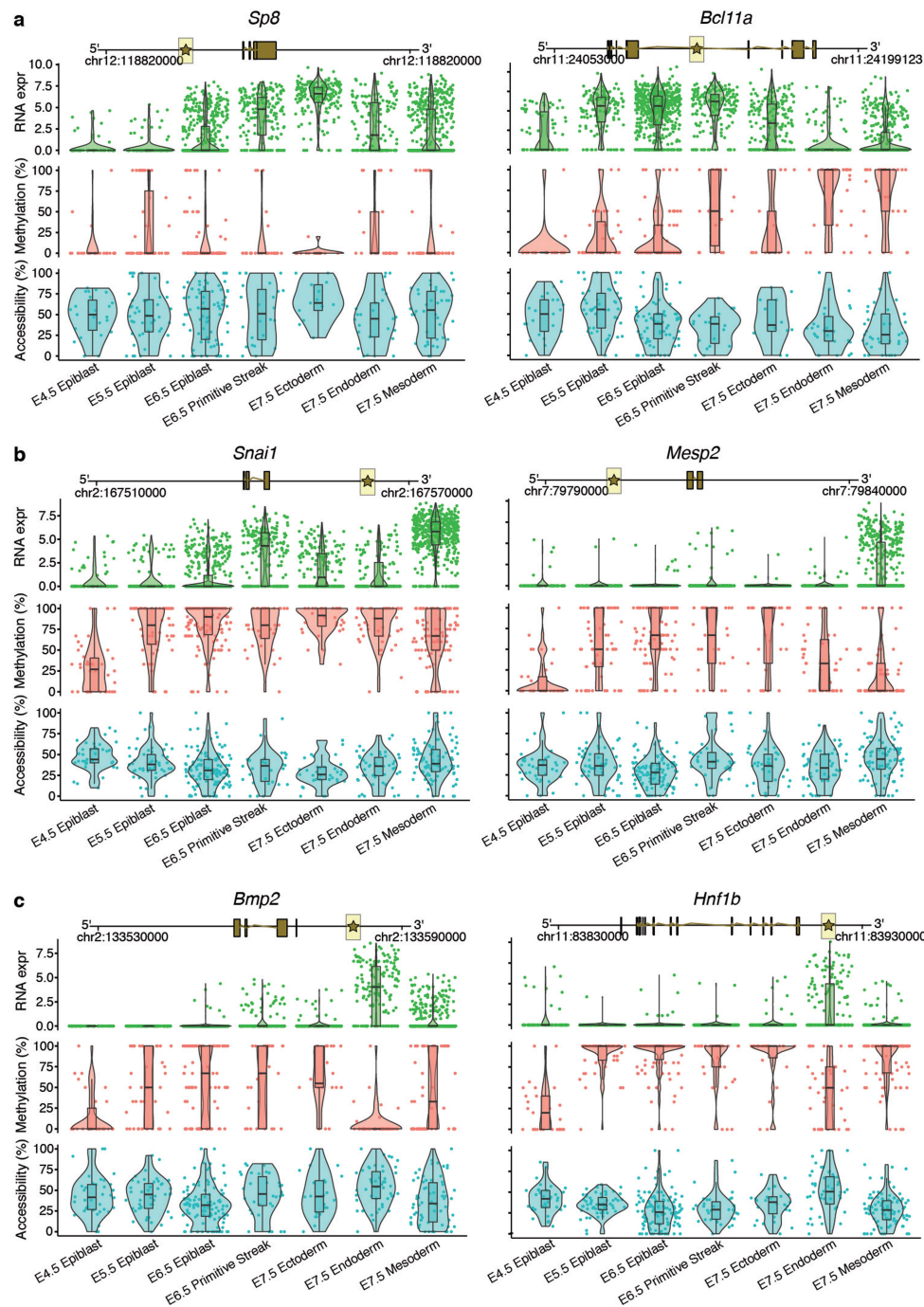
Extended Data Fig. 5 | Characterization of lineage-specific H3K27ac and H3K4me3 ChIP-seq data. **a**, Percentage of peaks overlapping promoters (± 500 bp of TSS of annotated mRNAs (Ensembl v.87); lighter colour) and not overlapping promoters (distal peaks, darker colour). H3K27ac peaks tend to be distal from the promoters, marking putative enhancer elements⁵³. H3K4me3 peaks tend to overlap promoter regions, marking TSS⁵⁴. **b**, Venn diagrams showing overlap of peaks for each lineage, for distal H3K27ac (left) and H3K4me3 (right). This shows that H3K27ac peaks tend to be lineage-specific, whereas H3K4me3 peaks tend to be shared between lineages. **c**, Illustrative example of the ChIP-seq profile for the ectoderm marker *Cxcl12*. The top tracks show wiggle plots of ChIP-seq read density (normalized by total read count)

for lineage-specific H3K27ac and H3K4me3. The coding sequence is shown in black. The bottom tracks show the lineage-specific peak calls (Methods). H3K27ac peaks are split into distal (putative enhancers) and proximal to the promoter. **d**, Left, bar plot of the fraction of E7.5 lineage-specific enhancers ($n = 691$ for ectoderm, 618 for endoderm and 340 for mesoderm) that are uniquely marked by H3K27ac in either E10.5 midbrain, E12.5 gut or E10.5 heart. Right, heat map displaying H3K27ac levels at individual lineage-specific enhancers ($n = 2,039$ for ectoderm, 1,124 for endoderm and 631 for mesoderm) in more differentiated tissues. E7.5 enhancers are predominantly marked in their differentiated-tissue counterparts (midbrain for ectoderm, gut for endoderm and heart for mesoderm).



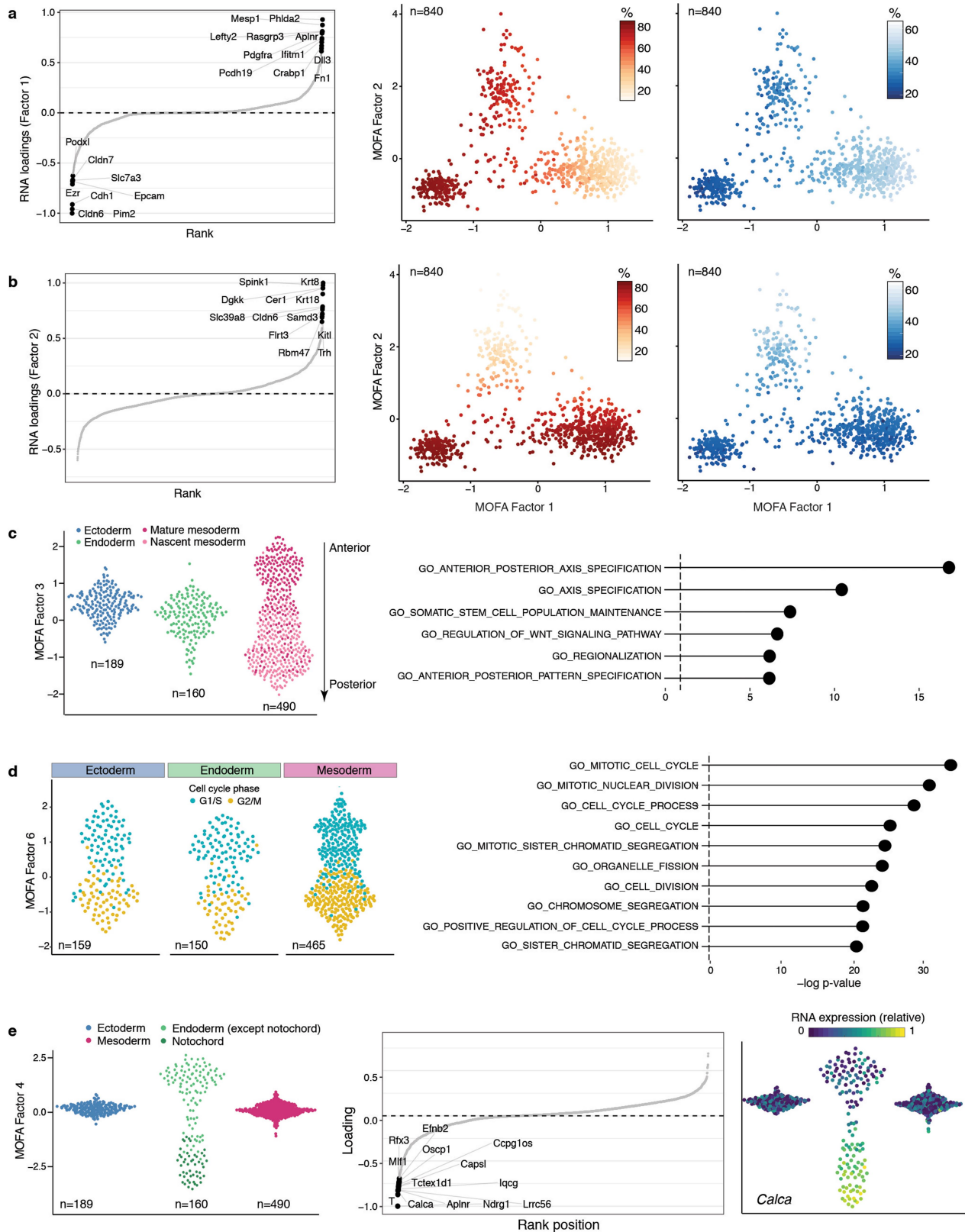
Extended Data Fig. 6 | Differential DNA methylation and chromatin accessibility analysis at E7.5 for different genomic contexts. a. Bar plots showing the fraction (left) or the total number (right) of differentially methylated (red) or accessible (blue) loci ($FDR < 10\%$, y axis) per genomic context (x axis). Each subplot corresponds to the comparison of one cell type (group A) against cells comprising the other cell types present at E7.5 (group B). In the graphs on the right, positive values indicate an increase in DNA methylation or chromatin accessibility in group A, whereas negative values indicate a decrease in DNA methylation or chromatin accessibility. Differential

analysis of DNA methylation and chromatin accessibility was performed independently for each genomic element using a two-sided Fisher's exact test of equal proportions (Methods). **b.** Scatter plots showing differential DNA methylation (x axis) versus chromatin accessibility (y axis) analysis at promoters. Ectoderm versus non-ectoderm cells (left), endoderm versus non-endoderm cells (middle) and mesoderm versus non-mesoderm cells (right) are shown. Each dot corresponds to a gene ($n = 2,038$). Labeled black dots highlight genes with lineage-specific RNA expression that show significant differential methylation or accessibility in their promoters ($FDR < 10\%$).



Extended Data Fig. 7 | Illustrative examples of putative epigenetic regulation in enhancer elements during germ-layer commitment. a–c, Box and violin plots showing the distribution of RNA expression (\log_2 counts, green), enhancer DNA methylation (red) and chromatin accessibility (blue) levels for key germ-layer markers per stage and cell type. Marker genes for ectoderm (a), mesoderm (b) and endoderm (c) are shown. Box plots show

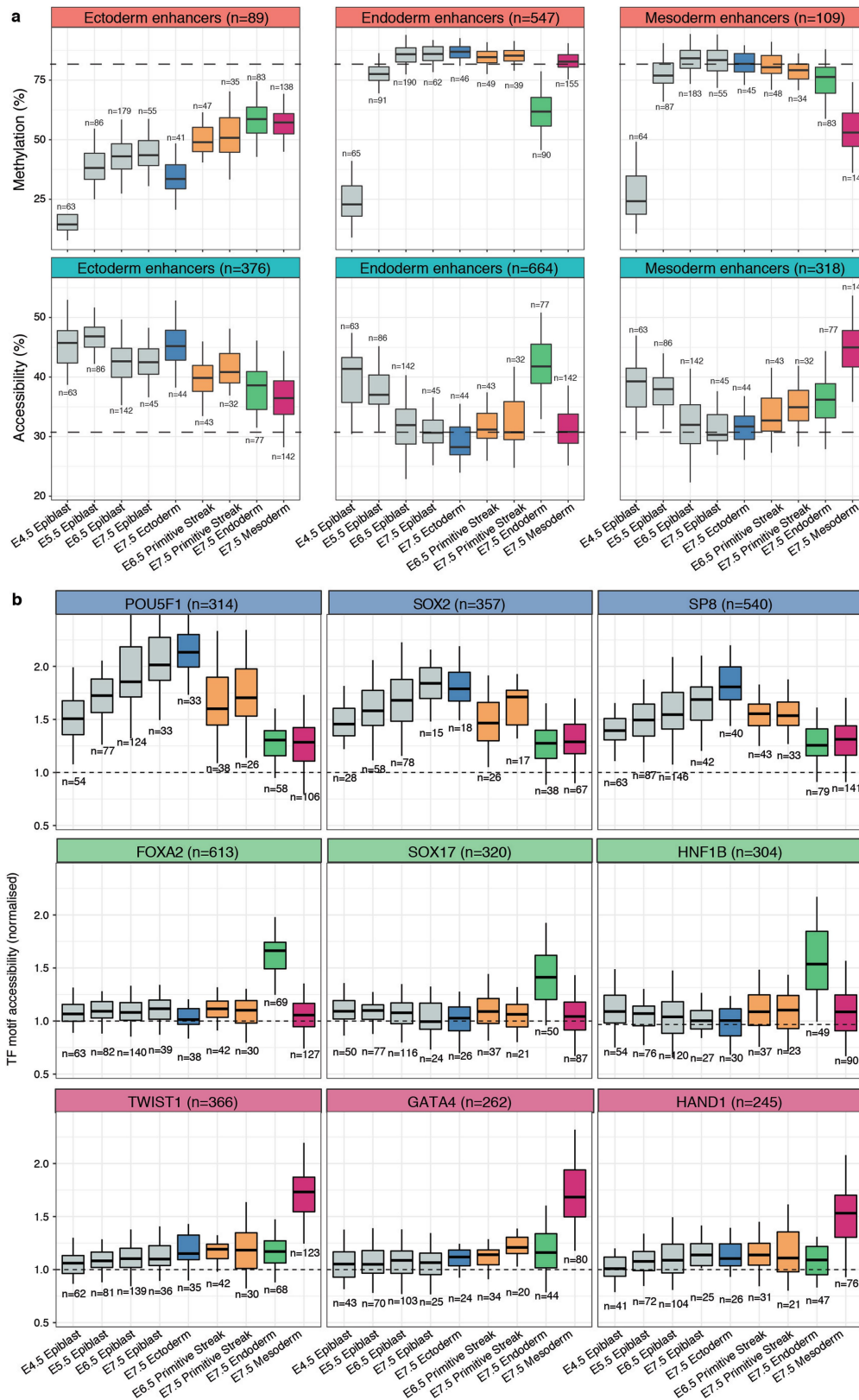
median levels and the first and third quartile, whiskers show $1.5 \times$ the interquartile range. Each dot corresponds to a single cell. For each gene, a genomic track is shown on the top. The enhancer region that is used to quantify DNA methylation and chromatin accessibility levels is represented with a star and highlighted in yellow. Genes were linked to putative enhancers by overlapping genomic coordinates with a maximum distance of 50 kb.



Extended Data Fig. 8 | See next page for caption.

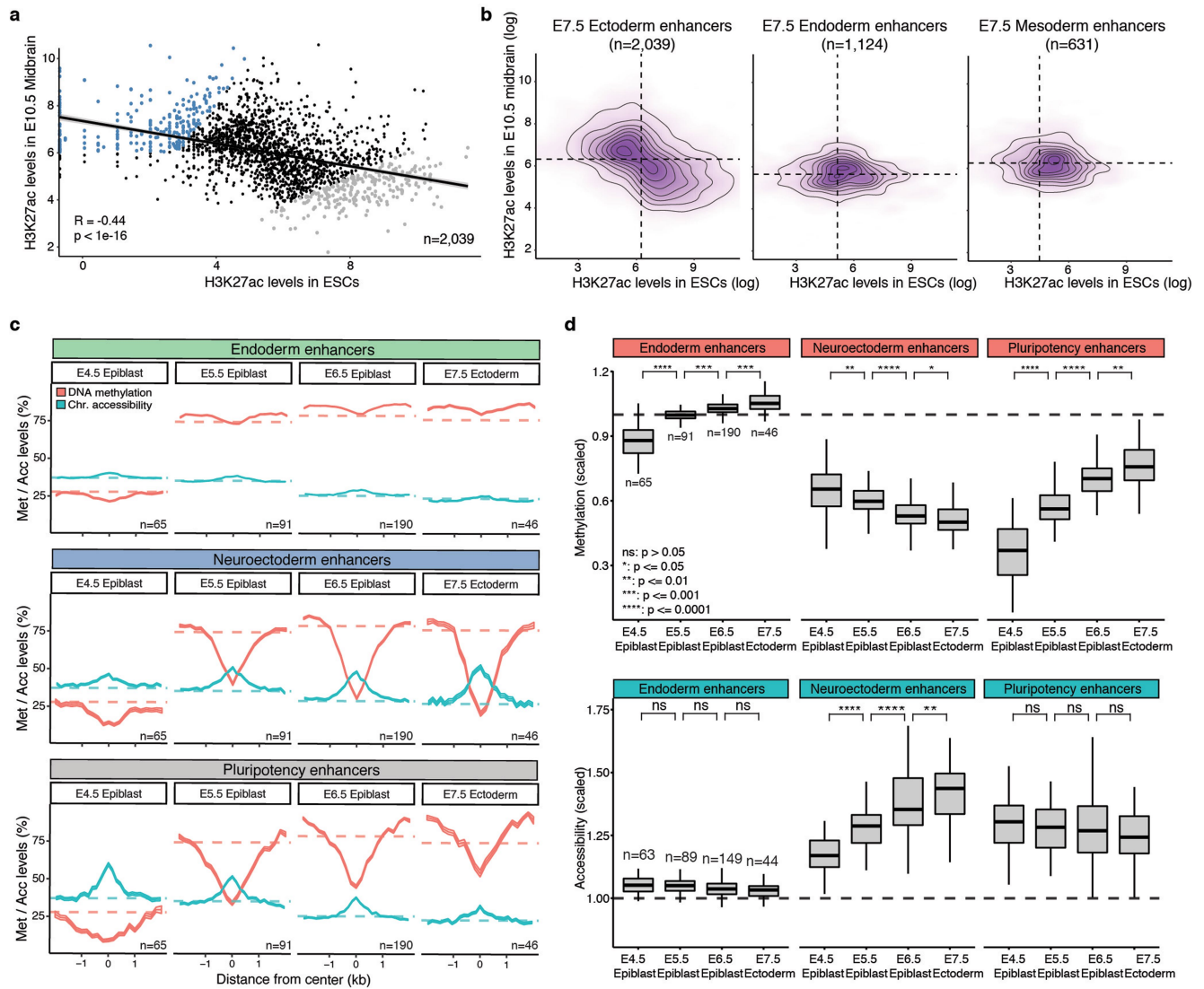
Extended Data Fig. 8 | Characterization of MOFA factors. **a**, Factor 1 as mesoderm commitment factor. Left, RNA-expression loadings for factor 1. Genes with large positive loadings increase expression in the positive factor values (mesoderm cells). Middle, scatter plot of factor 1 (x axis) and factor 2 (y axis) values. Each dot corresponds to a single cell, coloured by the average methylation levels of the top 100 enhancers with highest loading. Right, as the middle panel, except cells are coloured by the average accessibility levels. **b**, Factor 2 as the endoderm commitment factor. Left, RNA-expression loadings for factor 2. Genes with large positive loadings increase expression in the positive factor values (endoderm cells). Middle, scatter plot of factor 1 (x axis) and factor 2 (y axis) values. Each dot corresponds to a single cell, coloured by the average methylation levels (%) of the top 100 enhancers with highest loading. Right, as the middle panel, but cells are coloured by the average accessibility levels. **c**, Characterization of MOFA factor 3 as anteroposterior axial patterning and mesoderm maturation. Left, bee swarm plot of factor 3 values, grouped and coloured by cell type. The mesoderm cells are

subclassified into nascent and mature mesoderm (Extended Data Fig. 2). Right, gene set enrichment analysis of the gene loadings of factor 3. The top most significant pathways from MSigDB C2⁵⁵ (Methods) are shown. **d**, Characterization of MOFA Factor 6 as cell cycle. Left, bee swarm plot of factor 6 values, grouped by cell type and coloured by inferred cell-cycle state using cyclone⁵⁶ (G1/2, cyan; G2/M, yellow). Right, gene set enrichment analysis of the gene loadings of factor 6. The top most significant pathways from MSigDB C2⁵⁵ are shown. **e**, Characterization of MOFA factor 4 as notochord formation. Left, bee swarm plot of factor 4 values, grouped and coloured by cell type. The endoderm cells are subclassified into notochord (dark green) and not notochord (green) (Extended Data Fig. 2). Middle, RNA-expression loadings for factor 4. Genes with large negative loadings increase expression in the negative factor values (notochord cells). Right, same bee swarm plots as in left but coloured by the relative RNA expression of *Calca* (gene with the highest loading).



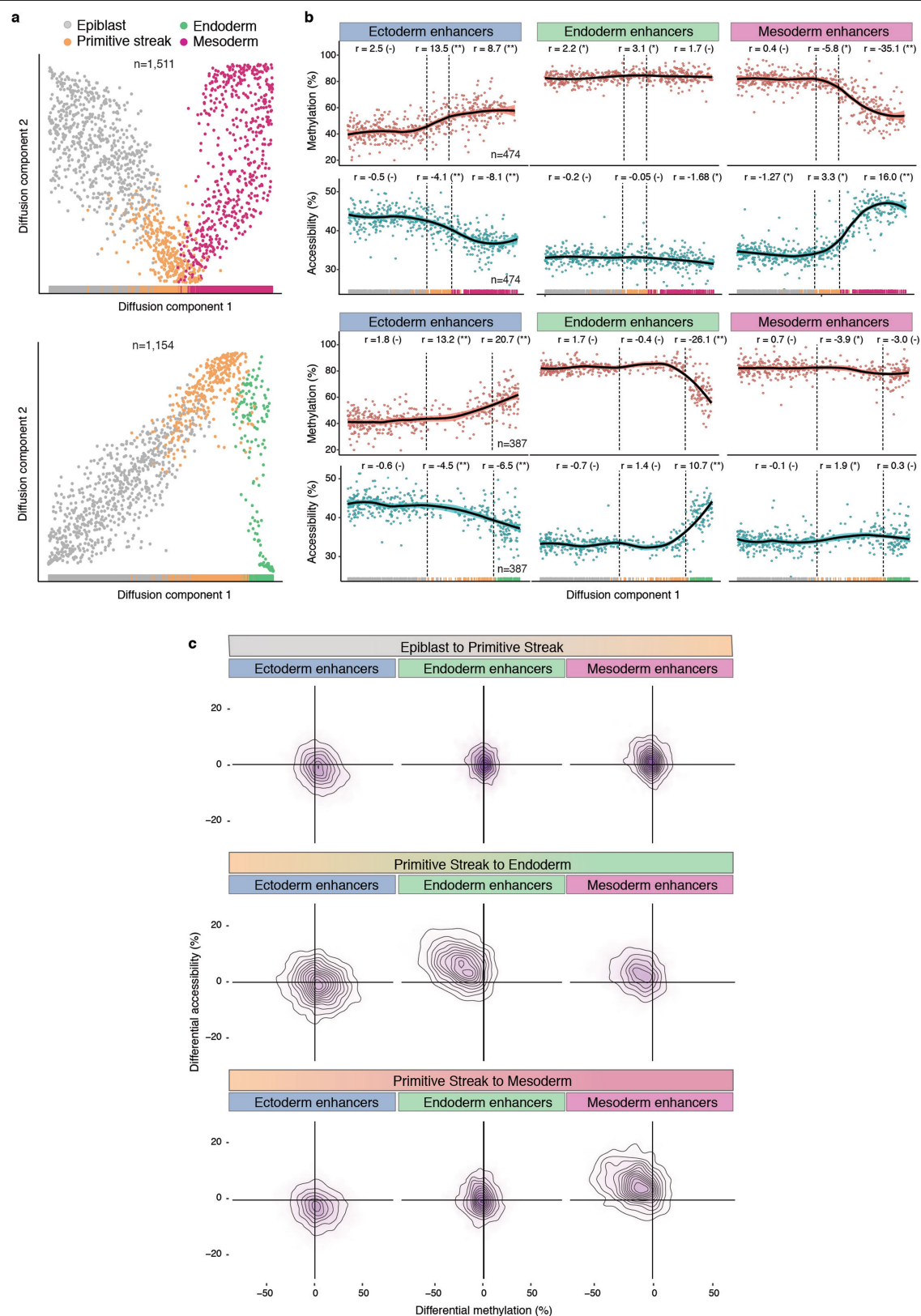
Extended Data Fig. 9 | DNA methylation and chromatin accessibility dynamics of E7.5 lineage-specific enhancers and transcription factor motifs across development. a. Box plots showing the distribution of DNA methylation (top) or chromatin accessibility (bottom) levels of E7.5 lineage-defining enhancers, across stages and cell types. Box plots show median levels and the first and third quartile, whiskers show 1.5× the interquartile range. The dashed lines represent the global background levels of DNA methylation at E7.5

(Extended Data Fig. 3). **b.** Box plots showing the distribution of chromatin accessibility levels (scaled to the genome-wide background) for 200-bp windows around transcription factor motifs associated with commitment to ectoderm (top), endoderm (middle) and mesoderm (bottom). Box plots show median levels and the first and third quartile, whiskers show 1.5× the interquartile range.



Extended Data Fig. 10 | E7.5 ectoderm enhancers contain a mixture of pluripotency and neural signatures with different epigenetic dynamics. a, Scatter plot showing H3K27ac levels for individual ectoderm enhancers ($n=2,039$) quantified in serum-grown ES cells (pluripotency enhancers, x axis) versus E10.5 midbrain (neuroectoderm enhancers, y axis). H3K27ac levels in the two lineages are negatively correlated (Pearson's $R = -0.44$), indicating that most enhancers are either marked in ES cells or in the brain. The top 250 enhancers that show the strongest differential H3K27ac levels between midbrain and ES cells (blue for midbrain-specific enhancers and grey for ES cell-specific enhancers) are highlighted. **b**, Density plots of H3K27ac levels in ES cells versus E10.5 midbrain. H3K27ac levels are negatively correlated at E7.5 ectoderm enhancers, but not in E7.5 endoderm ($n=1,124$) or mesoderm enhancers ($n=631$). **c**, Profiles of DNA methylation (red) and chromatin accessibility (blue) along the epiblast-ectoderm trajectory. Panels show different genomic contexts: E7.5 ectoderm enhancers that are specifically marked by H3K27ac in the midbrain (middle) or ES cells (bottom) (highlighted

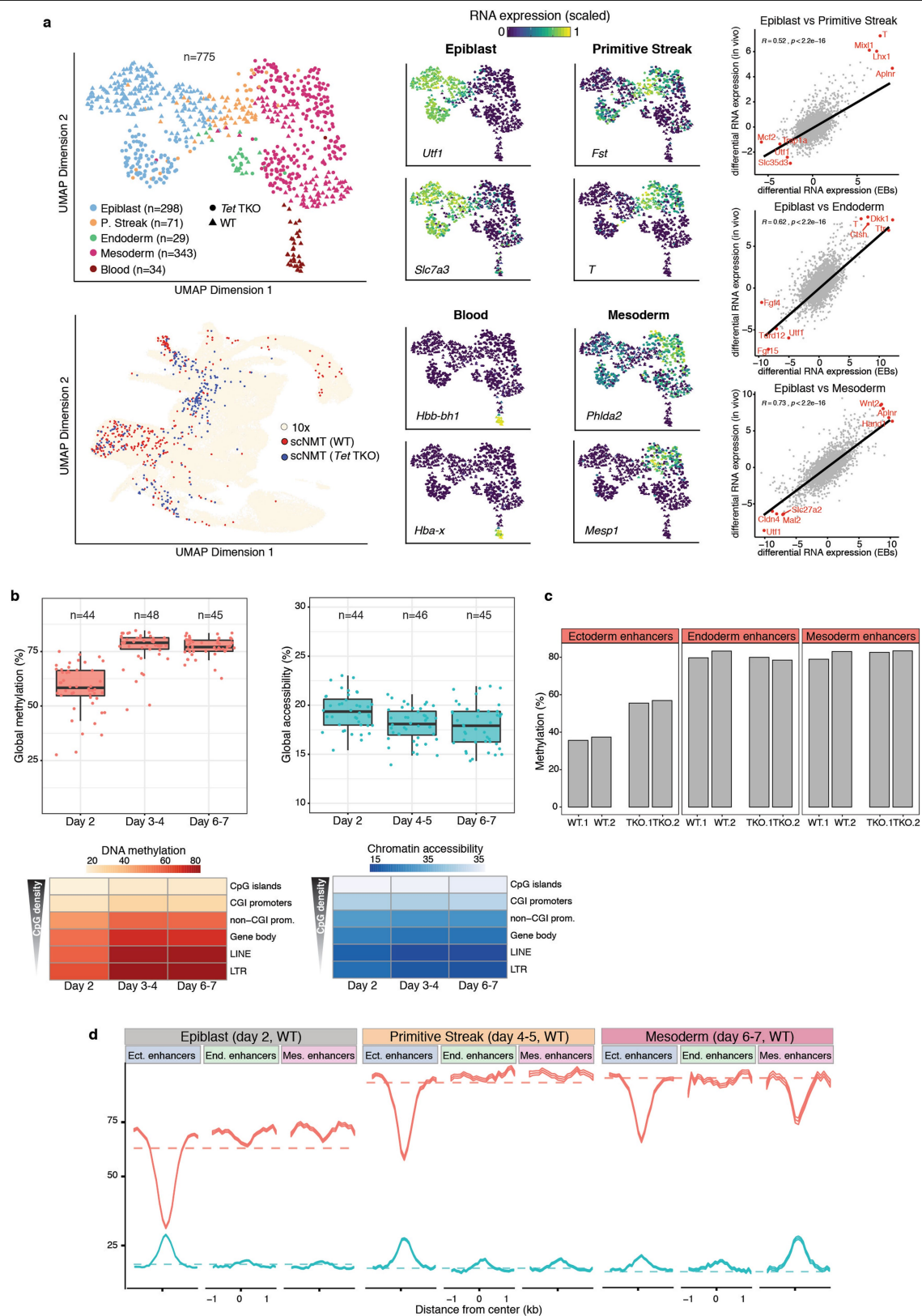
populations in **a**). Running averages of 50-bp windows around the centre of the ChIP-seq peaks (2 kb upstream and downstream) are shown. Solid lines display the mean across cells (within a given lineage) and shading displays the s.d. Dashed horizontal lines represent genome-wide background levels for DNA methylation (red) and chromatin accessibility (blue). For comparison, we have also incorporated E7.5 endoderm enhancers (top), which follow the genome-wide repressive dynamics. **d**, Box plots of the distribution of DNA methylation (top) and chromatin accessibility (bottom) levels along the epiblast-ectoderm trajectory. Panels show different genomic contexts: E7.5 ectoderm enhancers that are specifically marked by H3K27ac in the midbrain (middle) or ES cells (right) (highlighted populations in **a**). Box plots show median levels and the first and third quartile, whiskers show $1.5 \times$ the interquartile range. Dashed lines denote background DNA methylation and chromatin accessibility levels at the corresponding stage and lineage. For comparison, we have also incorporated E7.5 endoderm enhancers (left), which follow the genome-wide repressive dynamics.



Extended Data Fig. 11 | See next page for caption.

Extended Data Fig. 11 | Silencing of ectoderm enhancers precedes activation of mesoderm and endoderm enhancers. **a**, Reconstructed mesoderm (top) and endoderm (bottom) commitment trajectories using a diffusion pseudotime method applied to the RNA-expression data (Methods). Scatter plots of the first two diffusion components are shown, with cells coloured according to their lineage assignment ($n = 1,154$ for endoderm and $n = 1,511$ for mesoderm). For both cases, ranks along the first diffusion component are selected to order cells according to their differentiation state. **b**, DNA methylation (red) and chromatin accessibility (blue) dynamics of lineage-defining enhancers along the mesoderm (top) and endoderm (bottom)

trajectories. Each dot denotes a single cell ($n = 387$ for endoderm and $n = 474$ for mesoderm) and black curves represent non-parametric locally estimated scatterplot smoothing regression estimates. In addition, for each scenario we fit a piecewise linear regression model for epiblast, primitive streak and mesoderm or endoderm cells (vertical lines indicate the discretized lineage transitions). For each model fit, the slope (r) and its significance level are displayed in the top (– for nonsignificant, $0.01 < P < 0.1$ and $**P < 0.01$). **c**, Density plots showing differential DNA methylation (x axis) and chromatin accessibility (y axis) at lineage-defining enhancers calculated for each of the lineage transitions.



Extended Data Fig. 12 | See next page for caption.

Article

Extended Data Fig. 12 | Embryoid bodies recapitulate the transcriptional, methylation and accessibility dynamics of the embryo. a, Embryoid bodies show high transcriptional similarity to gastrulation-stage embryos. Top left, UMAP projection of RNA expression for the embryoid body dataset ($n = 775$). Cells are coloured by lineage assignment and shaped by genotype (WT or *Tet* TKO). Bottom left, UMAP projection of stages E6.5 to E8.5 of the atlas dataset (no extra-embryonic cells) with the nearest neighbours that were used to assign cell type labels to the scNMT-seq embryoid body dataset coloured in red (WT) or blue (*Tet* TKO). Middle, UMAP projection of embryoid body cells coloured by the relative RNA expression of marker genes. Right, scatter plot of the differential gene expression (\log_2 normalized counts) between different assigned lineages for embryoid bodies (x axis) versus embryos (y axis). Each dot represents one gene. Pearson correlation coefficient with corresponding *P* value (two-sided) are displayed. Lines show the linear regression fit. The top-four genes with the largest differential expression are highlighted in red. **b,** Global DNA methylation and chromatin accessibility levels during embryoid body differentiation. Top, box plots showing the distribution of genome-wide

CpG methylation (left) or GpC accessibility levels (right) per time point and lineage (compare with Extended Data Fig. 3). Each dot represents a single cell (only wild-type cells are used). Box plots show median levels and the first and third quartile, whiskers show $1.5 \times$ the interquartile range. Bottom, heat map of DNA methylation (left) or chromatin accessibility (right) levels per time point and genomic context (compare with Fig. 1e, f). **c,** Ectoderm enhancers are more methylated in *Tet* TKO compared with wild-type epiblast cells in vivo. Bar plots show the mean (bulk) DNA methylation levels for ectoderm (left), endoderm (middle) and mesoderm (right) enhancers in E6.5 epiblast cells²⁵. For each genotype, two replicates are shown. **d,** Profiles of DNA methylation (red) and chromatin accessibility (blue) at lineage-defining enhancers quantified over different lineages across embryoid body differentiation (only wild-type cells). Running averages in 50-bp windows around the centre of the ChIP-seq peaks (2 kb upstream and downstream) are shown. Solid lines display the mean across cells and shading displays the corresponding s.d. Dashed horizontal lines represent genome-wide background levels for methylation (red) and accessibility (blue).

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | n/a | Confirmed |
|-------------------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Sequencing was performed using an Illumina Nextseq500 instrument running NextSeq Control Software v4.0

Data analysis

All analysis code is available at https://github.com/rargelaguet/scnmt_gastrulation

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw sequencing data together with processed files (RNA counts, CpG methylation reports, GpC accessibility reports) are available in the Gene Expression Omnibus under accession GSE121708. A link to the processed data is available in the GitHub project.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were determined in order to obtain at least 50 cells for each germ layer.
Data exclusions	Regions of coverage outliers and extreme strand bias excluded as these were assumed to be alignment artefacts.
Replication	For each developmental stage, we collected cells from at least 3 individual embryos and results were consistent across embryos.
Randomization	This is not relevant since we did not use different experimental groups or conditions in our study.
Blinding	This is not relevant since we did not use different experimental groups or conditions in our study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	Tet ^l l ^{-/-} , 2 ^{-/-} , 3 ^{-/-} (C57BL/6J129/FVB) and matching wild-type mouse ES cells (Hu, X. et al. Cell Stem Cell 2014)
Authentication	None
Mycoplasma contamination	All cell lines tested negative for mycoplasma contamination with the MycoAlert testing kit (Lonza).
Commonly misidentified lines (See ICLAC register)	None

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Mus musculus, C57BL/6J. Embryos at 4.5 to 7.5 days post fertilization. Sex was unknown at the time of collection due to early embryonic stage.
Wild animals	Study did not involve wild animals.
Field-collected samples	Study did not involve field-collected samples.
Ethics oversight	All mouse experiments were approved by the Babraham Institute Animal Welfare and Ethical Review Body.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Cryo-EM structures of apo and antagonist-bound human Ca_v3.1

<https://doi.org/10.1038/s41586-019-1801-3>

Received: 23 July 2019

Accepted: 1 November 2019

Published online: 25 November 2019

Yanyu Zhao^{1,2,3,7}, Gaoxingyu Huang^{3,7}, Qiurong Wu^{3,7}, Kun Wu^{4,7}, Ruiqi Li⁴, Jianlin Lei⁵, Xiaojing Pan³ & Nieng Yan^{3,6*}

Among the ten subtypes of mammalian voltage-gated calcium (Ca_v) channels, Ca_v3.1–Ca_v3.3 constitute the T-type, or the low-voltage-activated, subfamily, the abnormal activities of which are associated with epilepsy, psychiatric disorders and pain^{1–5}. Here we report the cryo-electron microscopy structures of human Ca_v3.1 alone and in complex with a highly Ca_v3-selective blocker, Z944^{6,7}, at resolutions of 3.3 Å and 3.1 Å, respectively. The arch-shaped Z944 molecule reclines in the central cavity of the pore domain, with the wide end inserting into the fenestration on the interface between repeats II and III, and the narrow end hanging above the intracellular gate like a plug. The structures provide the framework for comparative investigation of the distinct channel properties of different Ca_v subfamilies.

Ca_v channels mediate Ca²⁺ influx into the cytosol in response to changes in membrane potential—a vital process that translates electrical signals on the membrane to chemical signals within the cell. Because of the broad spectrum of signalling events that involve Ca²⁺, Ca_v channels participate in a variety of physiological processes, such as contraction, secretion, gene expression and cell death. Ten genes encoding the pore-forming subunits of mammalian Ca_v channels have been identified, and the products are divided into three classes: Ca_v1 (Ca_v1.1–Ca_v1.4), Ca_v2 (Ca_v2.1–Ca_v2.3) and Ca_v3 (Ca_v3.1–Ca_v3.3) (Extended Data Fig. 1)^{1–4}.

The Ca_v3 channels only share around 20% sequence identity and around 45% similarity with the other two subfamilies (Extended Data Fig. 1a, b and Supplementary Fig. 1). Cloning of a Ca_v3 channel, human Ca_v3.1, was achieved 12 years after that of rabbit Ca_v1.1, the first Ca_v channel to be cloned^{8,9}. Sequence variations spread throughout the entire sequence, including the selectivity filter (SF). While Ca_v1 and Ca_v2 members all have four Glu residues (EEEE)—one on the corresponding locus of each repeat—that define the Ca²⁺ selectivity, the corresponding loci in the last two repeats are replaced by Asp in Ca_v3 channels (EEDD)^{9–12}. Furthermore, characterization of recombinantly expressed channels supports autonomous function of Ca_v3 core subunits, whereas the other families require auxiliary subunits for proper membrane localization and activity modulation^{5,13}.

The Ca_v3 channels can be activated at low membrane potentials—even lower than resting potential^{14–16}—and a hyperpolarization is often required for deinactivation^{13,14,17}. Thus, Ca_v3 members are designated as the low-voltage-activated (LVA) channels, whereas Ca_v1 and Ca_v2 are high-voltage-activated (HVA) channels^{18,19}. Ca_v3 channels are also known as T-type for transient single-channel conductance, in contrast to the larger, long-lasting conductance that is mediated by the L-type Ca_v1 channels^{2,3}. The unique properties of low-voltage activation, faster inactivation and slower deactivation of T-type voltage-gated calcium channels (VGCCs) support their

physiological role in cellular excitability modulation, low-threshold firing and pacemaker activity¹³.

Mutations in T-type VGCCs have been identified as associated with epilepsy seizures, ataxia and psychiatric disorders, and Ca_v3.2 knock-out mice show attenuated pain responses⁵ (Extended Data Fig. 1a and Extended Data Table 1). Development of T-type VGCC-selective antagonists thus represents a tempting strategy to mitigate these debilitating conditions²⁰. Z944, a highly Ca_v3-selective blocker, is a drug candidate in phase II clinical trials for the treatment of seizures and neuropathic pain. It blocks T-type channels, with a half maximal inhibitory concentration (IC₅₀) ranging between 50 and 160 nM, and shows 260–2,000-fold selectivity over HVA Ca²⁺ channels, Na_v1.5 and human ether-a-go-go-related gene (hERG) K⁺ channels^{6,7}. The specific binding site and the molecular basis for subtype selectivity remain unclear.

Cryo-electron microscopy (cryo-EM) structures of the tissue-extracted rabbit Ca_v1.1 complex containing the core subunit α1 and the auxiliary subunits β_{1a}, α2δ-1 and γ have been solved^{21,22}. However, the endogenous source for sample preparation prevented structural investigation of Ca_v mutants with pathophysiological significance and mechanistic implications. Structural elucidation of a recombinantly expressed Ca_v channel will be invaluable for the establishment of the structure–function relationship.

Despite previous success with recombinant Na_v channels^{23–26}, Ca_v channels have defied our extensive efforts to obtain a sufficient amount of proteins suitable for structural analysis until a proper construct was identified. A splice variant of rat Ca_v3.1 with a deletion within the I–II linker, designated Ca_v3.1-Δ8b, was shown to increase surface expression²⁷. We engineered a corresponding human Ca_v3.1-Δ8b variant, in which residues 509–642 were deleted from the full-length protein (UniProt ID, O43497-9) (Extended Data Fig. 1c and Supplementary Fig. 1). Cryo-EM structures were solved for human Ca_v3.1-Δ8b alone and in complex with Z944 at overall resolutions of 3.3 Å and 3.1 Å, respectively.

¹Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou, China. ²Institute of Biology, Westlake Institute for Advanced Study, Hangzhou, China. ³State Key Laboratory of Membrane Biology, Beijing Advanced Innovation Center for Structural Biology, Tsinghua-Peking Joint Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing, China. ⁴Medical Research Center, Beijing Key Laboratory of Cardiopulmonary Cerebral Resuscitation, Beijing Chao-Yang Hospital, Capital Medical University, Beijing, China. ⁵Technology Center for Protein Sciences, Ministry of Education Key Laboratory of Protein Sciences, School of Life Sciences, Tsinghua University, Beijing, China.

⁶Present address: Department of Molecular Biology, Princeton University, Princeton, NJ, USA. ⁷These authors contributed equally: Yanyu Zhao, Gaoxingyu Huang, Qiurong Wu, Kun Wu.

*e-mail: nyan@princeton.edu

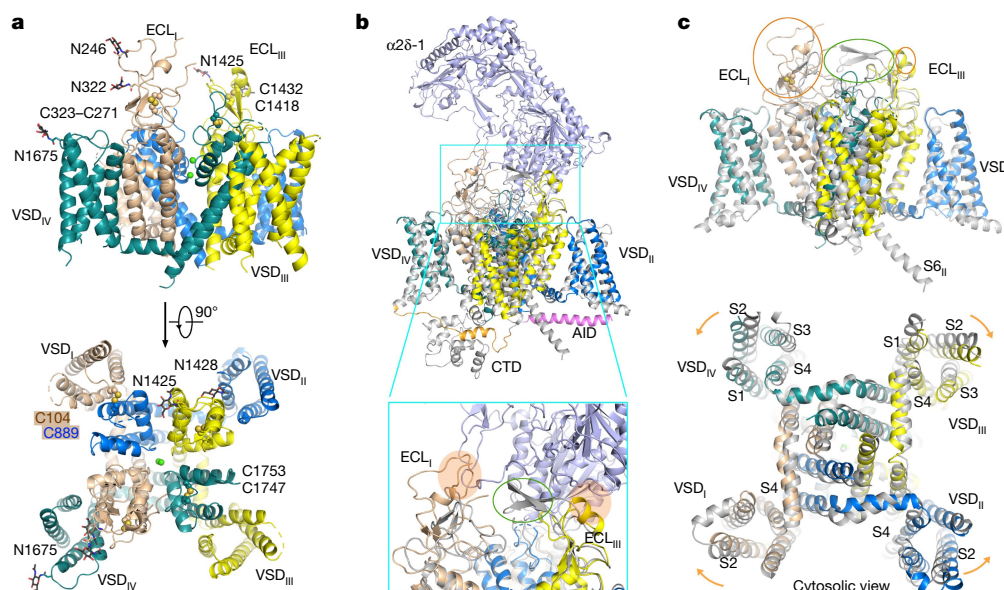


Fig. 1 | Structural differences between the L- and T-type VGCCs. a, Overall structure of human Ca_v3.1-Δ8b. The four repeats I–IV are coloured by domain. Disulfide bonds and sugar moieties are shown as spheres and black sticks, respectively. The disulfide bond between Cys104 and Cys889 is unique to T-type VGCCs. Two potential Ca²⁺ ions in the selectivity filter (SF) are shown as green spheres. ECL, extracellular loop. **b**, Structural basis for the incompatibility of α2δ association with the T-type Ca_v channels. Structures of Ca_v3.1 and nifedipine-bound Ca_v1.1 (PDB code 6JP5) are superimposed. For visual clarity, the β1 and γ subunits in the Ca_v1.1 complex are omitted. The α2δ-1 and α1 subunits of Ca_v1.1 are coloured light purple and grey, respectively. The

segments that correspond to the AID motif (pink), III–IV linker (orange) and CTD in Ca_v1.1-α1 are invisible in the Ca_v3.1 structure. The inset shows the structural basis for the incompatibility between Ca_v3.1 and α2δ-1. The semi-transparent ovals indicate the potential clashes between Ca_v3.1 and α2δ-1, and the green circle indicates the α2δ-1-interacting segments that are missing in Ca_v3.1. **c**, Conformational differences between Ca_v3.1 and Ca_v1.1. A side and a cytosolic view of the superimposed structures are shown. The orange and green circles correspond to those in the inset of **b**. The orange arrows indicate the slight rotations of the indicated VSDs from the positions in Ca_v1.1 to Ca_v3.1.

Structure of human Ca_v3.1-Δ8b

The channel properties of both full-length protein and Ca_v3.1-Δ8b were characterized using whole-cell patch clamp in HEK293T cells. Rat Ca_v3.1-Δ8b showed an increase of around two fold in conductance density, with activation and inactivation properties unchanged from the full-length form²⁷. Similarly, human Ca_v3.1-Δ8b exhibited a conductance density that was increased by around 1.5 fold, although both activation and steady-state inactivation curves are slightly left-shifted (Extended Data Fig. 2a and Extended Data Table 2).

After functional validation, we set out to resolve the structure of human Ca_v3.1-Δ8b. Details of protein generation, sample preparation and cryo-EM analysis are provided in Methods (see also Extended Data Fig. 2b–d). For simplicity, we will describe the Ca_v3.1-Δ8b structure as Ca_v3.1 and its complex with Z944 as the Z complex. Eventually, 105,559 and 138,449 selected particles gave rise to three-dimensional (3D) EM reconstructions at resolutions of 3.3 Å for the channel alone and 3.1 Å for the Z complex, respectively (Extended Data Figs. 2–4 and Extended Data Table 3).

The maps support reliable model building for most of the transmembrane and extracellular segments (Fig. 1a). None of the cytosolic segments observed in Ca_v1.1-α1 (Protein Data Bank (PDB) code 5GJV)²²—such as the α1-interacting domain (AID) helix, the elongated S6_{III}, the III–IV linker between S6_{III} and the voltage-sensing domain in the fourth repeat (VSD_{IV}), and the globular carboxy terminal domain—is visible in Ca_v3.1 (Fig. 1b and Supplementary Table 1).

The structure of Ca_v3.1 is consistent with that expected for an inactivated state that is characterized with depolarized or ‘up’ VSDs and closed intracellular gate (Extended Data Fig. 5). Extended extracellular segments between the pore-forming segment S5 and the pore helix P1 in repeats I and III, designated ECL_I and ECL_{III}, respectively, are resolved and each contains a short helix and a pair of anti-parallel

β-strands. The extracellular loops are stabilized by multiple disulfide bonds conserved in all Ca_v channels. An additional disulfide bond is structurally revealed between Cys104 on the S1–S2 linker in VSD_I and Cys889 on ECL_{II} (Fig. 1a and Supplementary Fig. 1). The unique disulfide bond, which stabilizes the interaction between VSD_I relative to the pore domain, may be responsible for the T-type-specific redox modulation of activation and inactivation kinetics²⁸.

α2δ and T-type VGCC structural incompatibility

Three conformations of the pore domain have been observed in the Ca_v1.1 structures, which we defined as classes I–III (PDB codes 5GJV, 6JP5 and 6JPA, respectively)²⁹. The structure of Ca_v3.1 can be superimposed to that of classes I and II Ca_v1.1, with a root-mean-square deviation (r.m.s.d.) of 1.97 Å over 705 Cα atoms and 2.17 Å over 765 Cα atoms, respectively (Fig. 1b). The VSDs, all in depolarized conformations in Ca_v1.1 and Ca_v3.1 (Extended Data Fig. 5), undergo positional shifts when the two structures are superimposed relative to the pore domain (Fig. 1c). Because of the relatively low sequence similarity (Supplementary Fig. 1), we refrain from overinterpreting the conformational differences between Ca_v3.1 and Ca_v1.1.

T-type VGCCs have been shown to function in the absence of auxiliary subunits. The intracellular segments that are required for binding to the β and γ subunits in Ca_v1.1 are invisible in the EM map of Ca_v3.1, precluding structure-based analysis. The resolved extracellular segments do provide an important clue to the incompatibility between T-type channels and α2δ, an extracellular subunit associated with the HVA channels (Fig. 1b, inset). A fragment of ECL_I and a short helix in ECL_{III} in Ca_v3.1 would clash with a loop region and the Cache1 domain in the α2δ subunit, respectively (Fig. 1b, inset). Although the loops may have the conformational flexibility to circumvent collision, the helix in ECL_{III} is rigidified through the disulfide bond between Cys1418

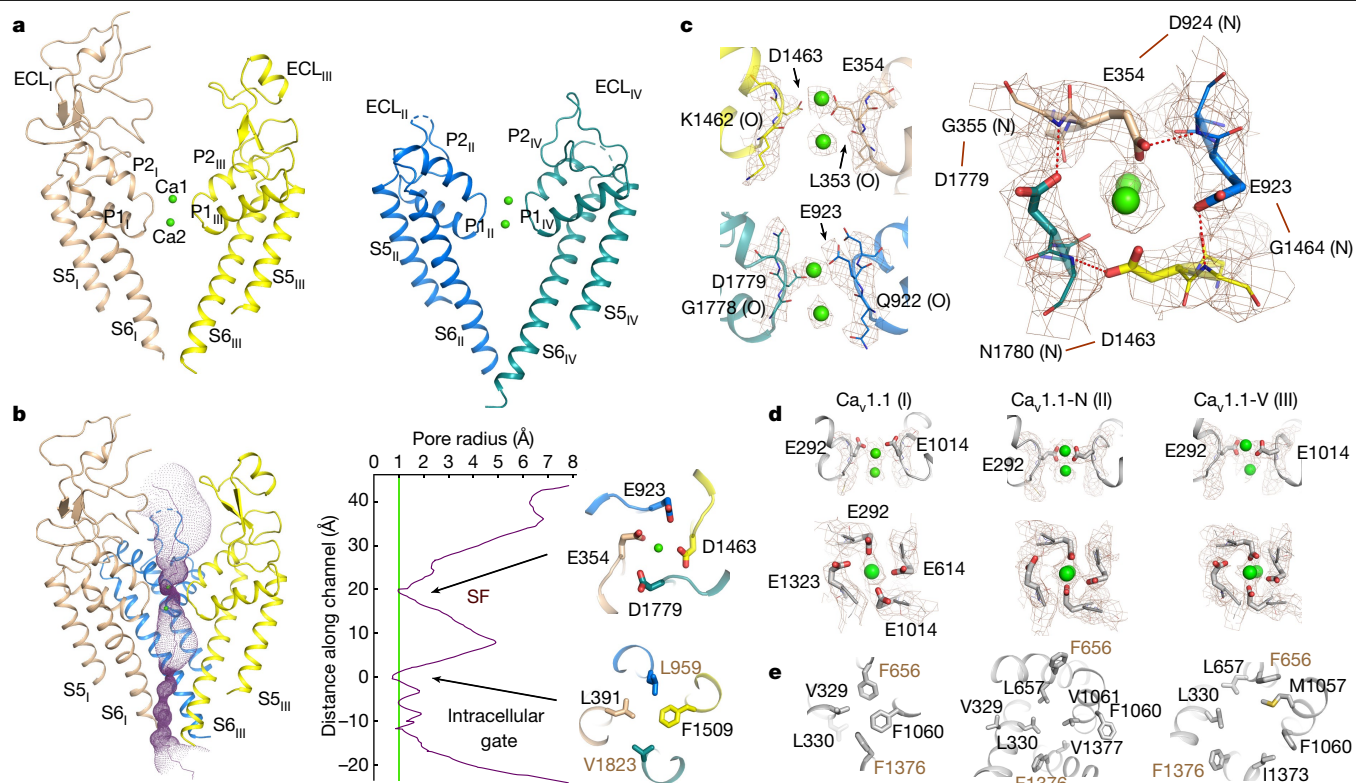


Fig. 2 | Selectivity filter and intracellular gate. **a**, Overall structure of the pore domain. Side views of the pore domain from the diagonal repeats are shown. Green spheres indicate two potential Ca^{2+} ions. **b**, The permeation path. The ion-conducting passage is illustrated by purple dots on the left, and the calculated pore radii are shown in the middle. Two constriction sites—the entrance to the SF enclosed by the EEDD motif and the intracellular gate—are shown on the right in extracellular views. **c**, The SF vestibule. Two side views (left) and an extracellular view (right) of the SF are shown. The densities for the residues that constitute the SF and the bound ions, shown as brown mesh, are

contoured at 5σ . The backbone carbonyl oxygen and amide are labelled as (O) and (N), respectively. **d**, The SF in the reported $\text{Ca}_v1.1$ structures. The densities are contoured at 4σ for $\text{Ca}_v1.1$ with 10 mM Ca^{2+} only ($\text{Ca}_v1.1$, class I), with nifedipine ($\text{Ca}_v1.1$ -N, class II) and with verapamil ($\text{Ca}_v1.1$ -V, class III). The EMDB codes for the three maps are EMD-9513, EMD-9866 and EMD-9868; the PDB codes are 5GJV, 6JP5 and 6JPA. **e**, Shifts of the intracellular gate in different $\text{Ca}_v1.1$ structures. Shown are extracellular views of the gate in the three $\text{Ca}_v1.1$ structures. Leu330/Phe656/Phe1060/Phe1376 in $\text{Ca}_v1.1$ correspond to Leu391/Leu959/Phe1509/Val1823 in $\text{Ca}_v3.1$. The varied residues are labelled brown.

and Cys1432, and thus is unlikely to avoid the conformational clash. In addition, a pair of anti-parallel β -strands in $\text{Ca}_v1.1$ that is responsible for the docking of the majority of the Cache1 domain in $\alpha 2\delta^{22}$ is missing in $\text{Ca}_v3.1$ (Fig. 1c, top). These missing or extra elements may collectively impede binding of $\alpha 2\delta$ to $\text{Ca}_v3.1$.

EEDD selectivity filter

The S5 and S6 helices from the four repeats constitute the pore domain (Fig. 2a, b). The intervening segments between S5 and S6 form the pore helices P1 and P2 that support the SF, with the construction site enclosed by Glu354/Glu923/Asp1463/Asp1779 (Fig. 2c). Densities that probably belong to two Ca^{2+} ions are resolved within the SF vestibule. The upper one is on the same plane as the carboxylate groups from the EEDD motif, and the lower one is caged by the C=O groups from the two preceding residues in each repeat (Fig. 2c and Extended Data Fig. 3c).

The side chains of the EEDD motif all exhibit similar conformations. The four carboxylate groups—each being stabilized by the amide of the residue that demarcates the SF loop and the P2 helix in the neighbouring repeat (Fig. 2c, right)—enclose a constriction site with a van der Waals diameter of approximately 2 Å (Fig. 2b). This site can only accommodate dehydrated Ca^{2+} ions. Notably, the constriction is executed by Glu354 and Asp1463 from repeats I and III, because the distance between the side groups of Glu923 and Asp1779 is approximately 2 Å longer (Fig. 2c, right).

The pore diameters estimated from the structure are smaller than the biophysical measurement of 5.1 Å for T-type and 6.2 Å for L-type

VGCCs based on the permeation ability of different organic cations³⁰. Rotation of the side chains of the EEEE or EEDD motif may enlarge the diameter of the filter for permeation of organic cations. Nevertheless, the nominal size of the SF constriction site in $\text{Ca}_v3.1$ appears smaller than that in $\text{Ca}_v1.1$ ²². A putative explanation is that the side chains of EEDD are further restricted through coordination by the neighbouring backbone amides (Fig. 2b–d).

The S6 tetrahelical bundle of $\text{Ca}_v3.1$ is sealed at the intracellular gate with three layers of hydrophobic residues along the permeation axis of the pore domain (Fig. 2b). The first constriction site beneath the central cavity comprises Leu391/Leu959/Phe1509/Val1823, which correspond to the gating residues Leu330/Phe656/Phe1060/Phe1376 in ligand-free $\text{Ca}_v1.1$ (Fig. 2e, left). Two additional layers of hydrophobic residues on the cytosolic side further secure gate closure (Fig. 2b).

Specific pore blockade by Z944

In the 3D EM reconstruction of the Z complex, the well-resolved Z944 adopts an arched shape and reclines in the cavity (Fig. 3a–c). The phenyl ring on one end projects into the fenestration enclosed by repeats II and III, and the tri-methyl group on the other end is positioned right above the intracellular gate (Fig. 3b, c). Z944 appears to combine the binding modes of both blockers and allosteric antagonists, the latter exemplified by the insertion of dihydropyridine to the III–IV fenestration²⁹. Binding to the II–III fenestration may underlie the state dependence for T-type inhibition by Z944⁶.

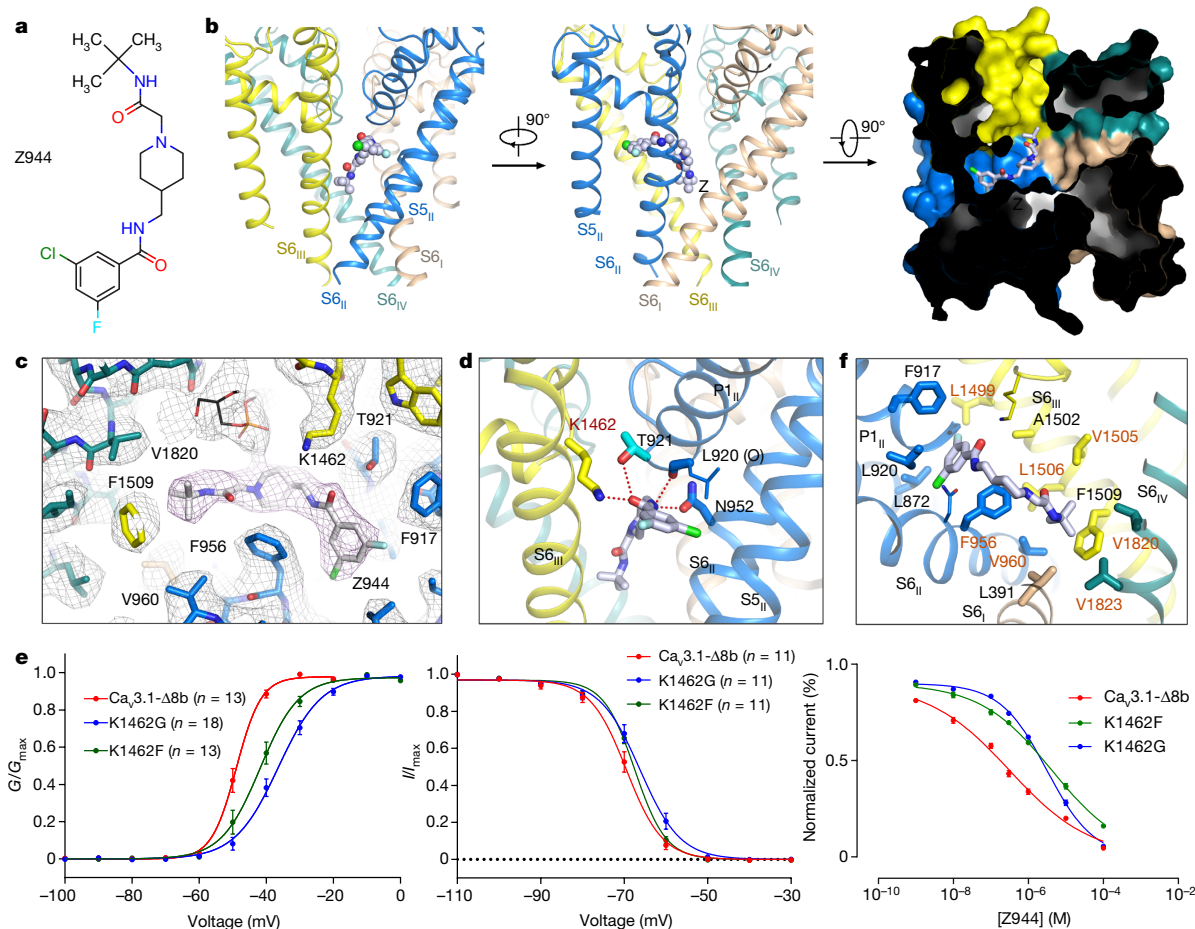


Fig. 3 | Specific blockade of T-type Ca_v channels by Z944. **a**, Chemical structure of Z944, generated in ChemDraw. **b**, Structural basis for pore blockade by Z944. A cut-open surface presentation viewed from the extracellular side is shown on the right. Z944, coloured silver, is shown either as spheres or sticks. **c**, EM map for Z944 and surrounding residues contoured at 7σ . A nearby lipid is shown as black sticks. **d**, Specific coordination of Z944 by polar residues. The potential electrostatic interactions are indicated by red dashed lines. **e**, Functional validation of the coordination of Z944 by the T-type specific Lys. The locus corresponding to Lys1462 is replaced by Phe in Ca_v1 and Gly in Ca_v2 channels. n values indicate the number of independent cells;

mean \pm s.e.m. Left and middle panels show voltage-dependent activation (left) and inactivation (right) curves of the indicated $\text{Ca}_v3.1\text{-}\Delta 8b$ variants. The right panel shows that single point mutation K1462F (green) or K1462G (blue) resulted in a change of the IC_{50} from 311 ± 25.6 nM to 4.2 ± 0.3 μM or 3.1 ± 0.2 μM , respectively. The sample sizes (n) tested from low to high concentrations are: $n = 4, 5, 5, 3, 8, 6, 3$ for $\text{Ca}_v3.1\text{-}\Delta 8b$; $n = 3, 4, 8, 8, 8, 8, 3$ for $\text{Ca}_v3.1\text{-}\Delta 8b$ (K1462F); and $n = 8, 8, 10, 10, 8, 6, 3$ for $\text{Ca}_v3.1\text{-}\Delta 8b$ (K1462G). G/G_{max} and I/I_{max} represent normalized conductance and ion current, respectively. **f**, Z944 is surrounded by hydrophobic residues on the S6 tetrahelical bundle. The residues that are not conserved in Ca_v1 and Ca_v2 channels are labelled in orange.

The phenyl ring is nearly perpendicular to the pore axis. The peptide bond between the phenyl ring and piperidine ring of Z944 is coordinated by several polar residues, including Thr921 on P1_{II} and Asn952 on S6_{II} , and the backbone $\text{C}=\text{O}$ of Leu920 on P1_{II} . The most prominent interaction is between the $\text{C}=\text{O}$ group of Z944 and the amine of Lys1462, the backbone of which constitutes the SF vestibule (Fig. 3d). Of particular note, Lys is unique to the T-type VGCCs at this locus, and is replaced by Phe in Ca_v1 and Gly in Ca_v2 channels (Supplementary Fig. 1). Therefore, this Lys residue may confer T-type specificity for Z944. Supporting this notion, substitution of Lys1462 in $\text{Ca}_v3.1\text{-}\Delta 8b$ with Phe or Gly, which causes a shift to the right of the activation curve by -7 mV and -12 mV, respectively, leads to decreased sensitivity to Z944, with IC_{50} increased from -0.3 μM to -4.2 μM and -3.1 μM , respectively (Fig. 3e, Extended Data Fig. 6a and Extended Data Table 2).

Fourteen hydrophobic residues on the P1_{II} helix and the S6 segments in repeats II–IV also participate in compound binding. S6_1 engages one residue, the gating residue Leu391, for interaction with one of the methyl groups on the narrow end of Z944 (Fig. 3f). Among these residues, only half are invariant in the other type of Ca_v channels (Fig. 3f and Supplementary Fig. 1). Although the varied residues are still hydrophobic, they may collectively change the contour and affinity

for compound accommodation, further contributing to subtype selectivity.

S6 shifts upon Z944 binding

Despite overall structural similarity with and without Z944, local structural shifts occur to the segments surrounding Z944 (Extended Data Fig. 6b, c). An evident consequence is the closure of the I–II fenestration (Fig. 4a). One helical turn on S6_{II} , comprising residues $_{951}\text{Gly-Asn-Tyr}_{953}$, undergoes an $\alpha \rightarrow \pi$ transition upon Z944 binding (Fig. 4b). Consequently, the ensuing fragment in S6_{II} undergoes an axial rotation by approximately one-third helical turn (Fig. 4c). The intracellular gate remains closed, but the gating residue on S6, changes from Leu959 to Val960 (Figs. 2b and 4c). Similar structural rearrangements have been observed in $\text{Ca}_v1.1$ ²⁹, the S6 helices of which in repeats I–III undergo $\alpha \rightarrow \pi$ transition in the middle upon binding to nifedipine or diltiazem while the gate remains closed.

Z944 binding also changes the distribution of transverse lipids in the pore domain (Extended Data Figs. 3b and 6d). An extra lipid molecule, the head of which points to the centre of the cavity, is found near Z944 (Fig. 3c). The density for the head group is contiguous with that in

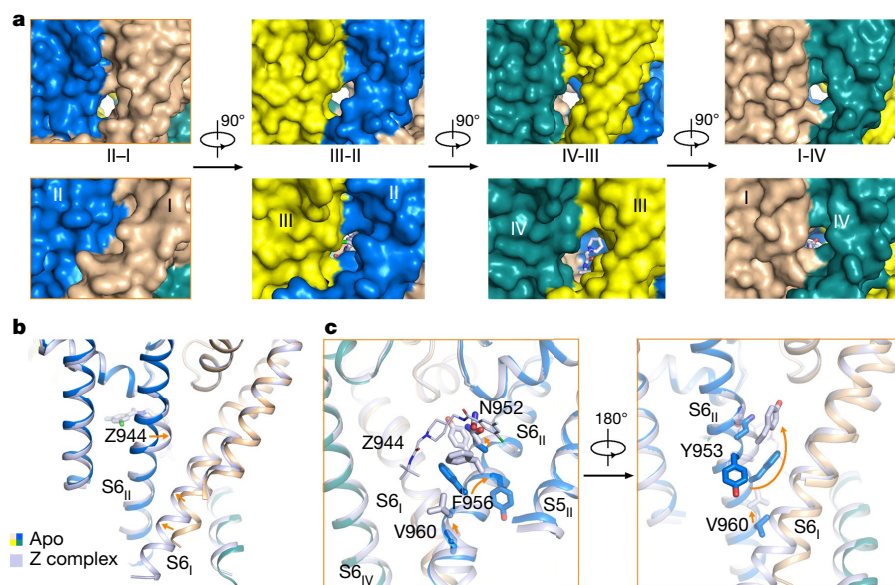


Fig. 4 | Local structural shifts upon Z944 binding. **a**, Closure of the fenestration on the I–II interface in the Z complex. The corresponding surface views of the four sides of the pore domain in the apo (upper row) and Z944-bound (lower row) structures are shown. **b**, Secondary structural transition in the middle of S6_{II} upon Z944 binding. The helical turn consisting of residues ₉₅₁Gly-Asn-Tyr₉₅₃ shifts from α to π helix upon Z944

binding, similar to that observed in Ca_v1.1²⁹. The structures of apo (coloured by domain) and Z944-bound (silver) Ca_v3.1 are superimposed. Orange arrows indicate the local shifts upon Z944 binding. **c**, Rotation of the bottom half of the S6_{II} helix upon Z944 binding. Consequently, Val960 replaces Leu959 to become the gating residue on S6_I.

the SF passage. Thus, whether the stretch of density in the SF indeed belongs to Ca²⁺ requires future investigation (Extended Data Fig. 6d, e).

Discussion

More than two dozen point mutations have been identified in Ca_v3.1 and Ca_v3.2 from patients with disorders such as epilepsy and spinocerebellar ataxia, among which 18 can be mapped to the resolved segments on Ca_v3.1 (Extended Data Fig. 7 and Extended Data Table 1). Structural determination of human Ca_v3.1, which was obtained through recombinant expression, not only provides the molecular basis for dissecting the pathogenic mechanism of these mutations but also establishes the framework for structural examination of various disease mutants.

Structural elucidation of representative members from the HVA and LVA families will facilitate mechanistic investigation of their distinct activation, inactivation and ion-conduction properties. We have captured the structures of Ca_v1.1 and Ca_v3.1 in three and two conformations, respectively. In the closed channels, the gate can be sealed by different residues as a result of the axial rotation of the S6 segments (Figs. 2e and 4c). These structural observations suggest that the channels may be more dynamic than observations using conductance states. Different conformations can support the same functional state. The presence of multi-layers of hydrophobic residues on the cytosolic edge of the S6 tetrahelical bundle may secure the closed state to avoid inadvertent leak. It is noted that two of the gating residues in Ca_v3.1, Leu959 and Val1823, have smaller side groups than the corresponding ones, Phe656 and Phe1376, in Ca_v1.1 (Fig. 2b, e). These variations may lower the energy for opening the intracellular gate, hence facilitating channel activation at low voltage. Further biophysical, computational and structural studies are required to elucidate the determinants for the mechanistic distinction between LVA and HVA channels and to establish the structure–function relationship.

The structure of Z944-bound Ca_v3.1 reveals the subtype-specific mode of action of a small molecule. Note that the III–IV fenestration of the L-type VGCCs and Na_v channels has been shown to accommodate dihydropyridine drugs and local anaesthetics, respectively^{29,31,32}.

The structure here reveals the II–III fenestration to be a specific drug-binding site. The fenestrations, owing to their relatively lower sequence conservation, thus represent specific druggable sites. However, the observed structural shifts upon ligand binding underscore the critical importance of structures of the complexes between target proteins and lead compounds for drug design. The structures presented here, and previously²⁹, lay out the framework for structure-guided drug discovery for the treatment of various Ca_v channelopathies.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-019-1801-3>.

- Clapham, D. E. Calcium signaling. *Cell* **131**, 1047–1058 (2007).
- Nowycky, M. C., Fox, A. P. & Tsien, R. W. Three types of neuronal calcium channel with different calcium agonist sensitivity. *Nature* **316**, 440–443 (1985).
- Ertel, E. A. et al. Nomenclature of voltage-gated calcium channels. *Neuron* **25**, 533–535 (2000).
- Zamponi, G. W., Striessnig, J., Koschak, A. & Dolphin, A. C. The physiology, pathology, and pharmacology of voltage-gated calcium channels and their future therapeutic potential. *Pharmacol. Rev.* **67**, 821–870 (2015).
- Dolphin, A. C. Voltage-gated calcium channels and their auxiliary subunits: physiology and pathophysiology and pharmacology. *J. Physiol.* **594**, 5369–5390 (2016).
- Tringham, E. et al. T-type calcium channel blockers that attenuate thalamic burst firing and suppress absence seizures. *Sci. Transl. Med.* **4**, 121ra19 (2012).
- Casillas-Espinosa, P. M. et al. Z944, a novel selective T-type calcium channel antagonist delays the progression of seizures in the amygdala kindling model. *PLoS One* **10**, e0130012 (2015).
- Tanabe, T. et al. Primary structure of the receptor for calcium channel blockers from skeletal muscle. *Nature* **328**, 313–318 (1987).
- Perez-Reyes, E. et al. Molecular characterization of a neuronal low-voltage-activated T-type calcium channel. *Nature* **391**, 896–900 (1998).
- Yang, J., Ellinor, P. T., Sather, W. A., Zhang, J. F. & Tsien, R. W. Molecular determinants of Ca²⁺ selectivity and ion permeation in L-type Ca²⁺ channels. *Nature* **366**, 158–161 (1993).
- Ellinor, P. T., Yang, J., Sather, W. A., Zhang, J. F. & Tsien, R. W. Ca²⁺ channel selectivity at a single locus for high-affinity Ca²⁺ interactions. *Neuron* **15**, 1121–1132 (1995).

12. Talavera, K. et al. Aspartate residues of the Glu-Glu-Asp-Asp (EEDD) pore locus control selectivity and permeation of the T-type Ca^{2+} channel α_{1G} . *J. Biol. Chem.* **276**, 45628–45635 (2001).
13. Perez-Reyes, E. Molecular physiology of low-voltage-activated T-type calcium channels. *Physiol. Rev.* **83**, 117–161 (2003).
14. Deschênes, M., Paradis, M., Roy, J. P. & Steriade, M. Electrophysiology of neurons of lateral thalamic nuclei in cat: resting properties and burst discharges. *J. Neurophysiol.* **51**, 1196–1219 (1984).
15. Zhan, X. J., Cox, C. L., Rinzel, J. & Sherman, S. M. Current clamp and modeling studies of low-threshold calcium spikes in cells of the cat's lateral geniculate nucleus. *J. Neurophysiol.* **81**, 2360–2373 (1999).
16. Aizenman, C. D. & Linden, D. J. Regulation of the rebound depolarization and spontaneous firing patterns of deep nuclear neurons in slices of rat cerebellum. *J. Neurophysiol.* **82**, 1697–1709 (1999).
17. Burlhis, T. M. & Aghajanian, G. K. Pacemaker potentials of serotonergic dorsal raphe neurons: contribution of a low-threshold Ca^{2+} conductance. *Synapse* **1**, 582–588 (1987).
18. Carbone, E. & Lux, H. D. A low voltage-activated, fully inactivating Ca channel in vertebrate sensory neurones. *Nature* **310**, 501–502 (1984).
19. Catterall, W. A. Structure and regulation of voltage-gated Ca^{2+} channels. *Annu. Rev. Cell Dev. Biol.* **16**, 521–555 (2000).
20. Zamponi, G. W. Targeting voltage-gated calcium channels in neurological and psychiatric diseases. *Nat. Rev. Drug Discov.* **15**, 19–34 (2016).
21. Wu, J. et al. Structure of the voltage-gated calcium channel $\text{Ca}_v1.1$ complex. *Science* **350**, aad2395 (2015).
22. Wu, J. et al. Structure of the voltage-gated calcium channel $\text{Ca}_v1.1$ at 3.6 Å resolution. *Nature* **537**, 191–196 (2016).
23. Shen, H. et al. Structure of a eukaryotic voltage-gated sodium channel at near-atomic resolution. *Science* **355**, eaal4326 (2017).
24. Shen, H., Liu, D., Wu, K., Lei, J. & Yan, N. Structures of human $\text{Na}_v1.7$ channel in complex with auxiliary subunits and animal toxins. *Science* **363**, 1303–1308 (2019).
25. Pan, X. et al. Structure of the human voltage-gated sodium channel $\text{Na}_v1.4$ in complex with $\beta 1$. *Science* **362**, eaau2486 (2018).
26. Pan, X. et al. Molecular basis for pore blockade of human Na^+ channel $\text{Na}_v1.2$ by the μ -conotoxin KIIIA. *Science* **363**, 1309–1313 (2019).
27. Shcheglovitov, A. et al. Alternative splicing within the I-II loop controls surface expression of T-type $\text{Ca}_v3.1$ calcium channels. *FEBS Lett.* **582**, 3765–3770 (2008).
28. Todorovic, S. M. et al. Redox modulation of T-type calcium channels in rat peripheral nociceptors. *Neuron* **31**, 75–85 (2001).
29. Zhao, Y. et al. Molecular basis for ligand modulation of a mammalian voltage-gated Ca^{2+} channel. *Cell* **177**, 1495–1506 (2019).
30. Cataldi, M., Perez-Reyes, E. & Tsien, R. W. Differences in apparent pore sizes of low and high voltage-activated Ca^{2+} channels. *J. Biol. Chem.* **277**, 45969–45976 (2002).
31. Lipkind, G. M. & Fozzard, H. A. Molecular modeling of local anesthetic drug binding by voltage-gated sodium channels. *Mol. Pharmacol.* **68**, 1611–1622 (2005).
32. Ahern, C. A., Eastwood, A. L., Dougherty, D. A. & Horn, R. Electrostatic contributions of aromatic residues in the local anesthetic receptor of voltage-gated sodium channels. *Circ. Res.* **102**, 86–94 (2008).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Article

Methods

Transient expression of human Ca_v3.1

Full-length cDNA of human Ca_v3.1 coding for 2,261 residues (Uniprot O43497-9) was a gift from J. Han (Xiamen University, China). For Ca_v3.1-Δ8b, residues 509–642 were deleted with standard two-step PCR. The full-length form and variants were cloned into the pCAG vector³³, with His₈ tag and Flag tag in tandem at the amino terminus. The constructs were verified by sequencing. For overexpression of Ca_v3.1-Δ8b, HEK293F cells (Invitrogen) were cultured in SMM 293T-I medium (Sino Biological) supplemented with 5% CO₂ in a Multitron-Pro shaker (Infors, 130 rpm) at 37 °C. When cell density reached 2 × 10⁶ cells per ml, 1.5 mg plasmids for Ca_v3.1-Δ8b were pre-incubated with 3.5 mg 25 kDa linear polyethylenimine (PEI) (Polysciences) in 50 ml fresh medium for 15–30 min for one litre of cell culture. The mixture was then added into cell culture to initiate the transfection. Transfected cells were cultured for 60–72 h before harvesting. No further authentication was performed for the commercially available cell line. Mycoplasma contamination was not tested.

Protein purification of human Ca_v3.1-Δ8b and complex preparation with Z944

For one batch of protein purification, 30 litres of transfected cells were harvested by centrifugation at 800g for 10 min and resuspended to 300 ml in lysis buffer containing 25 mM HEPES (pH 7.4), 100 mM NaCl, 25 mM KCl, 5 mM MgCl₂, and protease inhibitor cocktail containing 2 mM phenylmethylsulfonyl fluoride (PMSF), 2.6 μg ml⁻¹ aprotinin, 1.4 μg ml⁻¹ pepstatin and 10 μg ml⁻¹ leupeptin. After sonication on ice, the suspension was supplemented with *n*-dodecyl-β-D-maltopyranoside (DDM, Anatrace) to a final concentration of 1% (w/v), cholesteryl hemisuccinate Tris salt (CHS, Anatrace) to 0.2% (w/v), and ATP to 1 mM. After incubation at 4 °C for 2 h, the mixture was centrifuged at 150,000g for 30 min, and the supernatant was applied to 12 ml anti-Flag M2 affinity gel (Sigma) by gravity at 4 °C. The resin was washed four times, each with 10 ml buffer containing 25 mM HEPES (pH 7.4), 100 mM NaCl, 25 mM KCl, 5 mM MgCl₂, 0.02% (w/v) glyco-diosgenin (GDN, Anatrace), 1 mM ATP, and protease inhibitor cocktail for every 2 ml gel. The protein bound to 2 ml affinity gel was eluted with 10 ml buffer containing 25 mM Tris-Cl (pH 8.0), 100 mM NaCl, 25 mM KCl, 5 mM MgCl₂, 0.02% GDN, 1 mM ATP, 200 μg ml⁻¹ Flag peptide (Sigma), and protease inhibitor cocktail with gravity at 4 °C. The eluate was then applied to 8 ml Ni-NTA resin. The resin was washed three times, each with 10 ml buffer containing 25 mM Tris-Cl (pH 8.0), 100 mM NaCl, 25 mM KCl, 5 mM MgCl₂, 0.02% GDN, 1 mM ATP, 15 mM imidazole, and protease inhibitor cocktail for every 2 ml resin. The target protein was eluted via gravity at 4 °C, with 40 ml buffer containing 25 mM HEPES (pH 7.4), 150 mM NaCl, 10 mM CaCl₂, 0.02% GDN, 200 mM imidazole, and protease inhibitor cocktail. The eluate was then concentrated using a 100 kDa cut-off Centricon (Millipore) and further purified through size-exclusion chromatography (Superose-6, GE Healthcare). The peak fractions were stored at -80 °C for further experiments. A typical protein yield through this purification procedure was about 1–2 μg per litre cell culture. For cryo-sample preparation, target protein purified from -150 l cell culture was pooled and concentrated to 50 μl at a concentration of approximately 1 mg ml⁻¹.

To investigate the interaction between Ca_v3.1-Δ8b and different regulators, Z944 (Tocris) and calmodulin³⁴ were separately added to aliquoted Ca_v3.1-Δ8b at final concentrations of 100 μM and 150 μM, respectively. The mixtures were incubated at 4 °C for 30 min before being concentrated. The estimated protein concentration for each sample was approximately 1 mg ml⁻¹.

Whole cell electrophysiology

The HEK293T cells were cultured in Dulbecco's Modified Eagle Medium (DMEM, BI) containing 4.5 mg ml⁻¹ glucose and 10% fetal bovine serum (FBS, BI), and co-transfected with the expression plasmids for wild type

or mutations with an eGFP-encoding plasmid when cell confluency reached 70%. After incubation at 37 °C under 5% CO₂ for 24 h, cells were treated with 0.05% trypsin (BI) and put on poly-D-lysine (Sigma-Aldrich) coated 12-mm cover slips (Assistant) for electrophysiological characterization. All experiments were performed at room temperature. No further authentication was performed for the commercially available cell line. Mycoplasma contamination was not tested.

The whole-cell Ca²⁺ currents were recorded in HEK293T cells using an EPC-10 amplifier with Patchmaster 2.90.4 software (HEKA Elektronik) and glass micropipettes (2–3 MΩ, Sutter Instrument) made by P-97 pipette puller (Sutter Instrument). The electrodes were filled with the internal solution composed of (in mM) 130 CsCH₃SO₃, 10 TEA-Cl, 10 EGTA, 10 HEPES, 5 MgCl₂, 5 Na-ATP, pH 7.4 with CsOH, and the extracellular solution was composed of (in mM) 105 CsCl, 40 TEA-Cl, 2 CaCl₂, 1 MgCl₂, 10 D-glucose, 10 HEPES, pH 7.4 with CsOH. The data were analysed using Fitmaster 2.90.4 (HEKA Elektronik) and Prism 8.2.1 (GraphPad Software).

The holding potential in all experiments was -100 mV. The voltage dependence of ion current (*I*-*V*) was analysed using a protocol consisting of steps from a holding potential to voltages ranging from -100 to 70 mV for 150 ms in 10-mV increments. The linear component of leaky current and capacitive transients was subtracted using the -P/8 procedure. In the activation and conductance density calculation, we used the equation $G = I / (V - V_r)$, where V_r (the reversal potential) represents the voltage at which the current is zero. For the activation curves, the conductance (*G*) was normalized and plotted against the voltage from -100 mV to -20 mV or 0 mV. To obtain the conductance density curves, *G* was divided by the capacitance (*C*), and plotted against the voltage from -100 mV to -20 mV.

For voltage dependence of inactivation, cells were clamped at a holding potential, and were applied to step pre-pulses from -110 mV to -30 mV for 10 s with an increment of 10 mV. Then, the calcium current was recorded at the test pulse of -20 mV for 150 ms. The peak currents under the test pulses were normalized and plotted against the pre-pulse voltage. The time course of inactivation data from the peak current at -20 mV was fitted to a single exponential equation: $y = A1 \exp(-x/\tau_{\text{inac}}) + y0$, where *A1* is the relative fraction of current inactivation, τ_{inac} is the time constant, *x* is the time, and *y0* is the amplitude of the steady-state component. τ_{inac} values of Ca_v3.1-Δ8b, Ca_v3.1-Δ8b (K1462F) and Ca_v3.1-Δ8b (K1462G) were compared by one-way ANOVA and Tukey's test.

To investigate the state-dependent blockade of Ca_v3.1-Δ8b, Ca_v3.1-Δ8b (K1462F) and Ca_v3.1-Δ8b (K1462G) by Z944^{6,7,35–38}, the current was recorded after 50-ms pulses at 1 Hz from -100 mV to -20 mV. The compound solutions were perfused to the recording cell using a multichannel perfusion system (MPS-2, World Precision Instruments).

Sample sizes were chosen to give s.e.m. values of less than 10% of peak values based on experience with similar experiments. The GFP-positive cells were randomly selected for whole-cell patch clamp. All the constructs were recorded and analysed blindly to avoid bias.

Cryo-EM data acquisition

Aliquots of 3.5 μl of concentrated Ca_v3.1-Δ8b alone or in the presence of modulators were loaded onto glow-discharged holey carbon grids (Quantifoil Cu R1.2/1.3, 300 mesh). Grids were blotted for 2.0 s and plunge-frozen in liquid ethane cooled by liquid nitrogen using Vitrobot Mark IV (Thermo Fisher) at 8 °C and with 100% humidity. Grids were transferred to a Titan Krios electron microscope (Thermo Fisher) operating at 300 kV and equipped with a Gatan Gif Quantum energy filter (slit width 20 eV) and spherical aberration (Cs) image corrector. Micrographs were recorded using a K2 Summit counting camera (Gatan Company) in super-resolution mode with a nominal magnification of 105,000×, resulting in a calibrated pixel size of 0.545 Å. Each stack of 32 frames was exposed for 5.6 s, with an exposing time of 0.175 s per frame. The total dose for each stack was about 48 e⁻ per Å². AutoEMation was used for fully automated data collection³⁹. All 32 frames in

each stack were first aligned and summed using MotionCorr⁴⁰, with twofold binned to a pixel size of 1.091 Å per pixel. The output stacks from MotionCorr were further motion-corrected with MotionCor2⁴¹, and dose weighting was performed⁴². The defocus values were set from −1.3 to −1.8 µm and were estimated by Gctf⁴³.

Image processing

The protocol for image acquisition and processing is identical for different datasets. However, no extra density was observed for calmodulin after data processing reconstruction. It is not surprising as none of the cytosolic element has been resolved. We therefore combined the datasets with or without calmodulin for processing. The following description and Extended Data Fig. 2 refer to the final workflow for Ca_v3.1-Δ8b alone and in complex with Z944.

A total of 7,075 and 5,716 cryo-EM micrographs were collected, and 3,268,403 and 2,158,477 particles were auto-picked by RELION-2.1 for Ca_v3.1-Δ8b alone and in complex with Z944, respectively^{44,45}. Particle picking was performed using low-pass filtered templates to 20 Å to limit reference bias. All subsequent 2D and 3D classifications and refinements were performed using RELION-3.0-beta⁴⁶ or RELION-2.1. Multiple rounds of reference-free 2D classification using RELION-2.1 were performed to remove ice spots, contaminants and aggregates, yielding 1,271,083 and 1,165,545 particles for Ca_v3.1-Δ8b alone and in complex with Z944, respectively. The particles were processed with a global search $K = 1$ procedure using RELION-3.0-beta to determine the initial orientation alignment parameters using bin4 particles. A published EM map of human Na_v1.2 low-pass filtered to 100 Å was used as an initial reference²⁶. After 60 iterations, the datasets from the last four iterations were subject to local search multi-reference 3D classifications using 4–6 classes with an angular sampling step of 3.7° and searching range of 30°. The multi-reference models were generated using the reconstructions at the used iteration low-pass filtered to 8.8, 15, 25, 35, 45 and 55 Å, respectively. Lower-resolution references were removed if the class number was less than 6. Particles from good classes were then combined and re-extracted with a box size of 120 and binned pixel size of 2.182 Å for further 3D classification; 263,202 and 755,740 particles remained for the Ca_v3.1-Δ8b and Ca_v3.1-Δ8b in complex with Z944 datasets, respectively. Another round of multi-reference classification using bin2 particles yielded datasets containing 231,057 and 737,580 particles respectively, giving rise to reconstructions of Ca_v3.1-Δ8b alone and in complex with Z944, both at 4.4-Å resolution. The particles were then re-extracted using a box size of 320 and pixel size of 1.091 Å. Two additional rounds of multi-reference 3D classification and five rounds of random-phase 3D classification resulted in 105,559 and 138,449 particles that yielded respective reconstructions at 3.4 Å and 3.1 Å. Application of a core mask for further local-search refinement improved resolutions to 3.3 Å and 3.1 Å.

Reported resolutions are based on the gold-standard Fourier shell correlation (FSC) 0.143 criterion. Before visualization, all density maps were corrected for the modulation transfer function of the detector and sharpened by applying a negative B-factor that was estimated using automated procedures⁴⁷. Local resolution variations were estimated using RELION-2.1.

Model building and structure refinement

The starting model of Ca_v3.1-Δ8b was built in SWISS-MODEL⁴⁸ based on the structure of rabbit Ca_v1.1-α1 subunit (PDB 5GJV). The starting model was then manually docked into the 3.3-Å EM map in Chimera⁴⁹. The model was manually adjusted in COOT⁵⁰, followed by refinement against the corresponding maps by phenix.real_space_refine program in PHENIX⁵¹ with secondary structure and geometry restraints. A total of 984 residues were assigned with side chains in Ca_v3.1-Δ8b, and 5 sugar moieties and 9 lipid molecules were built. Intracellular regions were not modelled due to the lack of corresponding densities. For model building of the Z complex, the apo structure was docked into the map

initially. The restraint file of Z944 was generated using phenix.elbow in PHENIX. Then the protein model and Z944 molecule were also manually adjusted in COOT. Overfitting of the overall model was monitored by refining the model in one of the two half maps from the gold-standard refinement approach and testing the refined model against the other map⁵². Statistics of the map reconstruction and model refinement can be found in Extended Data Table 3. All structure figures were prepared in PyMol⁵³, and the ion-conducting passage was calculated by HOLE⁵⁴.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The atomic coordinates and EM maps for Ca_v3.1 alone and in complex with Z944 have been deposited in the PDB with the accession codes 6KZO and 6KZP, and the EMDb with the codes EMD-0791 and EMD-0792, respectively. Source Data for Fig. 3e and Extended Data Figs. 2a and 6a are available in the online version of the paper. All other data are available from the corresponding author upon reasonable request.

33. Matsuda, T. & Cepko, C. L. Electroporation and RNA interference in the rodent retina in vivo and in vitro. *Proc. Natl Acad. Sci. USA* **101**, 16–22 (2004).
34. Gong, D. et al. Modulation of cardiac ryanodine receptor 2 by calmodulin. *Nature* **572**, 347–351 (2019).
35. Lee, M. Z944: a first in class T-type calcium channel modulator for the treatment of pain. *J. Peripher. Nerv. Syst.* **19**, S11–S12 (2014).
36. LeBlanc, B. W. et al. T-type calcium channel blocker Z944 restores cortical synchrony and thalamocortical connectivity in a rat model of neuropathic pain. *Pain* **157**, 255–263 (2016).
37. Nam, G. T-type calcium channel blockers: a patent review (2012–2018). *Expert Opin. Ther. Pat.* **28**, 883–901 (2018).
38. Marks, W. N. et al. The T-type calcium channel blocker Z944 reduces conditioned fear in Genetic Absence Epilepsy Rats from Strasbourg and the non-epileptic control strain. *Eur. J. Neurosci.* **50**, 3046–3059 (2019).
39. Lei, J. & Frank, J. Automated acquisition of cryo-electron micrographs for single particle reconstruction on an FEI Tecnai electron microscope. *J. Struct. Biol.* **150**, 69–80 (2005).
40. Li, X. et al. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods* **10**, 584–590 (2013).
41. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
42. Grant, T. & Grigorieff, N. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *eLife* **4**, e06980 (2015).
43. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
44. Kimanius, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5**, e18722 (2016).
45. Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
46. Zivanov, J. et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7**, e42166 (2018).
47. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).
48. Waterhouse, A. et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
49. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
50. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
51. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
52. Amunts, A. et al. Structure of the yeast mitochondrial large ribosomal subunit. *Science* **343**, 1485–1489 (2014).
53. DeLano, W. L. The PyMOL Molecular Graphics System. <http://www.pymol.org> (2002).
54. Smart, O. S., Neduevelil, J. G., Wang, X., Wallace, B. A. & Sansom, M. S. HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph.* **14**, 354–360, 376 (1996).
55. Larkin, M. A. et al. Clustal W and clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
56. Bourinet, E. & Zamponi, G. W. Block of voltage-gated calcium channels by peptide toxins. *Neuropharmacology* **127**, 109–115 (2017).
57. IUPHAR/BPS. Guide to Pharmacology. <https://www.guidetopharmacology.org/> (2019).
58. Weiss, N., Black, S. A., Bladen, C., Chen, L. & Zamponi, G. W. Surface expression and function of Cav3.2 T-type calcium channels are controlled by asparagine-linked glycosylation. *Pflügers Arch.* **465**, 1159–1170 (2013).
59. Lazniewska, J., Rzhetsky, Y., Zhang, F. X., Zamponi, G. W. & Weiss, N. Cooperative roles of glucose and asparagine-linked glycosylation in T-type calcium channel expression. *Pflügers Arch.* **468**, 1837–1851 (2016).

60. Liu, Y. et al. Asparagine-linked glycosylation modifies voltage-dependent gating properties of Ca_v3.1-T-type Ca²⁺ channel. *J. Physiol. Sci.* **69**, 335–343 (2019).
61. Ondacova, K., Karmazinova, M., Lazniewska, J., Weiss, N. & Lacinova, L. Modulation of Cav3.2 T-type calcium channel permeability by asparagine-linked glycosylation. *Channels* **10**, 175–184 (2016).
62. Weiss, N., Black, S. A. G., Bladen, C., Chen, L. & Zamponi, G. W. Surface expression and function of Ca_v3.2 T-type calcium channels are controlled by asparagine-linked glycosylation. *Pflugers Arch. Eur. J. Physiol.* **465**, 1159–1170 (2013).
63. Jiang, Y. et al. X-ray structure of a voltage-dependent K⁺ channel. *Nature* **423**, 33–41 (2003).
64. Tao, X., Lee, A., Limapichat, W., Dougherty, D. A. & MacKinnon, R. A gating charge transfer center in voltage sensors. *Science* **328**, 67–73 (2010).
65. Coutelier, M. et al. A recurrent mutation in *CACNA1G* alters Cav3.1 T-type calcium-channel conduction and causes autosomal-dominant cerebellar ataxia. *Am. J. Hum. Genet.* **97**, 726–737 (2015).
66. Morino, H. et al. A mutation in the low voltage-gated calcium channel *CACNA1G* alters the physiological properties of the channel, causing spinocerebellar ataxia. *Mol. Brain* **8**, 89 (2015).
67. Chemin, J. et al. De novo mutation screening in childhood-onset cerebellar atrophy identifies gain-of-function mutations in the *CACNA1G* calcium channel gene. *Brain* **141**, 1998–2013 (2018).
68. Splawski, I. et al. *CACNA1H* mutations in autism spectrum disorders. *J. Biol. Chem.* **281**, 22085–22091 (2006).
69. Heron, S. E. et al. Extended spectrum of idiopathic generalized epilepsies associated with *CACNA1H* functional variants. *Ann. Neurol.* **62**, 560–568 (2007).
70. Chen, Y. et al. Association between genetic variation of *CACNA1H* and childhood absence epilepsy. *Ann. Neurol.* **54**, 239–243 (2003).
71. Heron, S. E. et al. Genetic variation of *CACNA1H* in idiopathic generalized epilepsy. *Ann. Neurol.* **55**, 595–596 (2004).
72. Meyer, K. et al. Mutations in disordered regions can cause disease by creating dileucine motifs. *Cell* **175**, 239–253 (2018).
73. Daniil, G. et al. *CACNA1H* mutations are associated with different forms of primary aldosteronism. *EBioMedicine* **13**, 225–236 (2016).
74. Scholl, U. I. et al. Recurrent gain of function mutation in calcium channel *CACNA1H* causes early-onset hypertension with primary aldosteronism. *eLife* **4**, e06315 (2015).

Acknowledgements We thank X. Li for technical support during EM image acquisition. We thank J. Han for sharing the cDNA for human Ca_v3.1 (Uniprot O43497-9). This work was funded by the National Natural Science Foundation of China (projects 81920108015, 31800628 and 31621092), and the National Key R&D Program (2016YFA0500402 to X.P. and 2016YFA0501100 to J.L.) from Ministry of Science and Technology of China. We thank the Tsinghua University Branch of China National Center for Protein Sciences (Beijing) for providing the cryo-EM facility support. We thank the computational facility support on the cluster of Bio-Computing Platform (Tsinghua University Branch of China National Center for Protein Sciences Beijing) and the ‘Explorer 100’ cluster system of Tsinghua National Laboratory for Information Science and Technology. N.Y. is supported by the Shirley M. Tilghman endowed professorship from Princeton University.

Author contributions N.Y. conceived the project. Y.Z. and Q.W. conducted molecular cloning and protein purification. Y.Z., G.H., Q.W. and J.L. performed experiments for structural determination. K.W., X.P. and R.L. performed and analysed electrophysiological measurements. All authors contributed to data analysis. N.Y. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

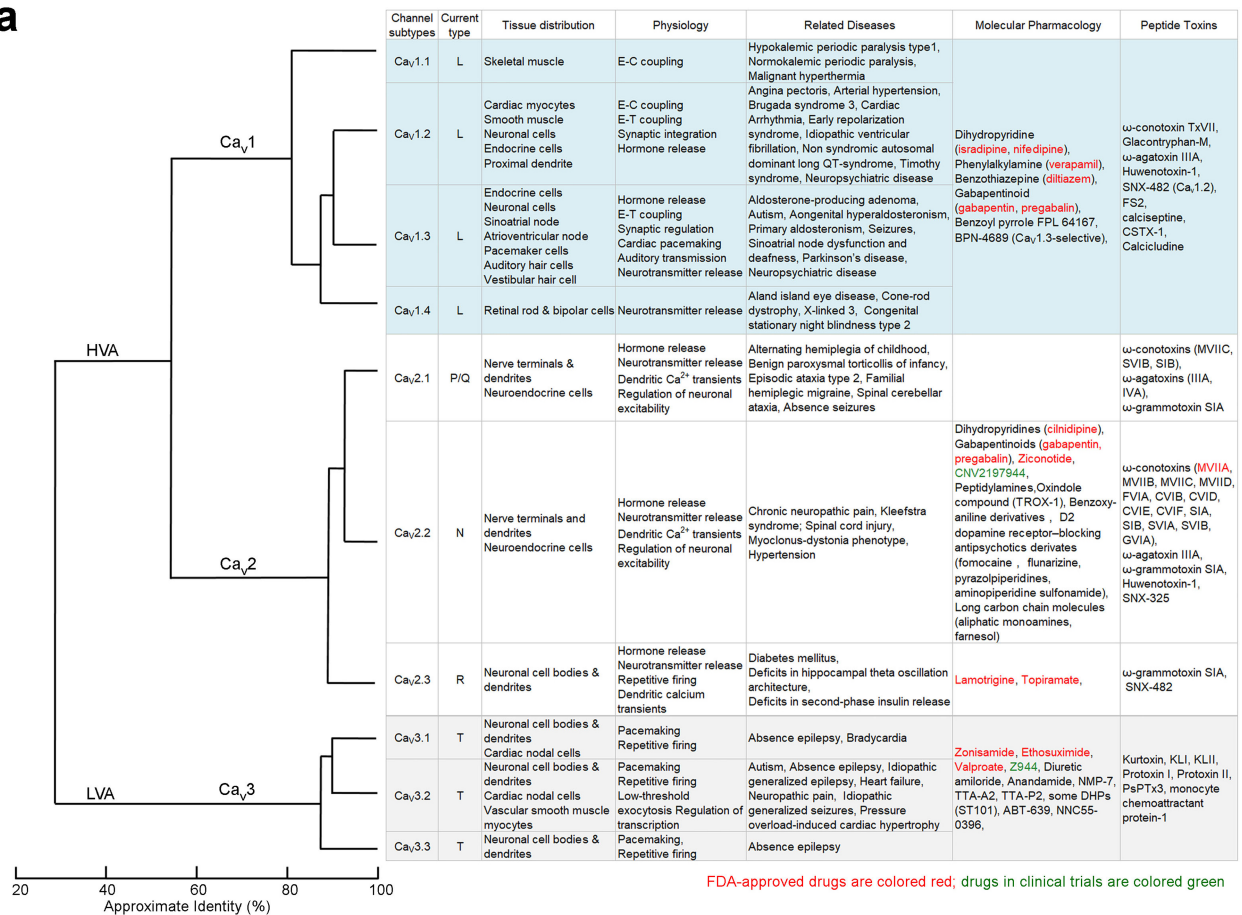
Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-019-1801-3>.

Correspondence and requests for materials should be addressed to N.Y.

Peer review information *Nature* thanks Jörg Striessnig, Gerald W. Zamponi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

a



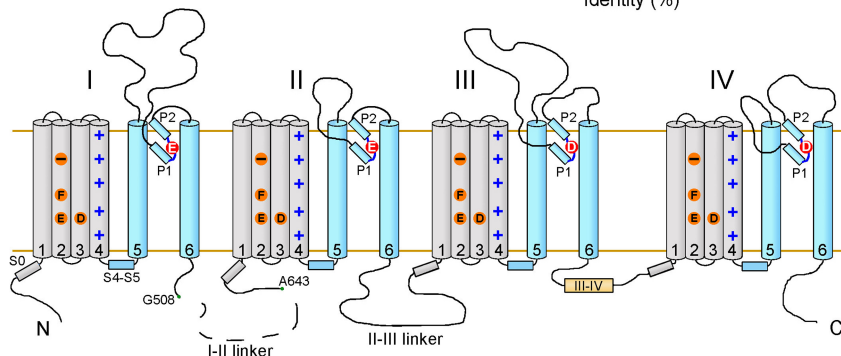
b

	Ca _v 1.1	Ca _v 1.2	Ca _v 1.3	Ca _v 1.4	Ca _v 2.1	Ca _v 2.2	Ca _v 2.3	Ca _v 3.1	Ca _v 3.2	Ca _v 3.3
Ca _v 1.1		71.5	72.9	54.86	52.8	54.9	55.5	44.5	45.0	47.0
Ca _v 1.2	55.0		79.9	72.1	54.1	56.2	55.8	44.0	44.1	45.1
Ca _v 1.3	55.1	63.7		76.0	55.2	56.3	55.5	43.8	45.7	45.5
Ca _v 1.4	54.5	55.0	59.9		52.5	57.6	55.9	46.4	46.9	46.9
Ca _v 2.1	31.3	32.0	32.0	30.8		76.5	71.2	45.1	43.3	46.5
Ca _v 2.2	32.9	33.3	33.4	34.5	59.4		73.4	46.3	45.3	46.7
Ca _v 2.3	32.6	32.5	31.8	32.4	54.0	54.3		46.1	46.4	48.0
Ca _v 3.1	19.8	19.8	19.9	20.8	19.7	21.2	20.0		67.3	65.6
Ca _v 3.2	20.3	19.3	20.2	20.5	18.5	20.3	20.1	51.2		68.3
Ca _v 3.3	21.6	20.8	20.2	21.9	20.0	21.1	20.2	48.6	51.1	

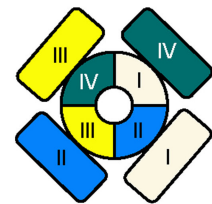
Identity (%)

Similarity (%)

c



d



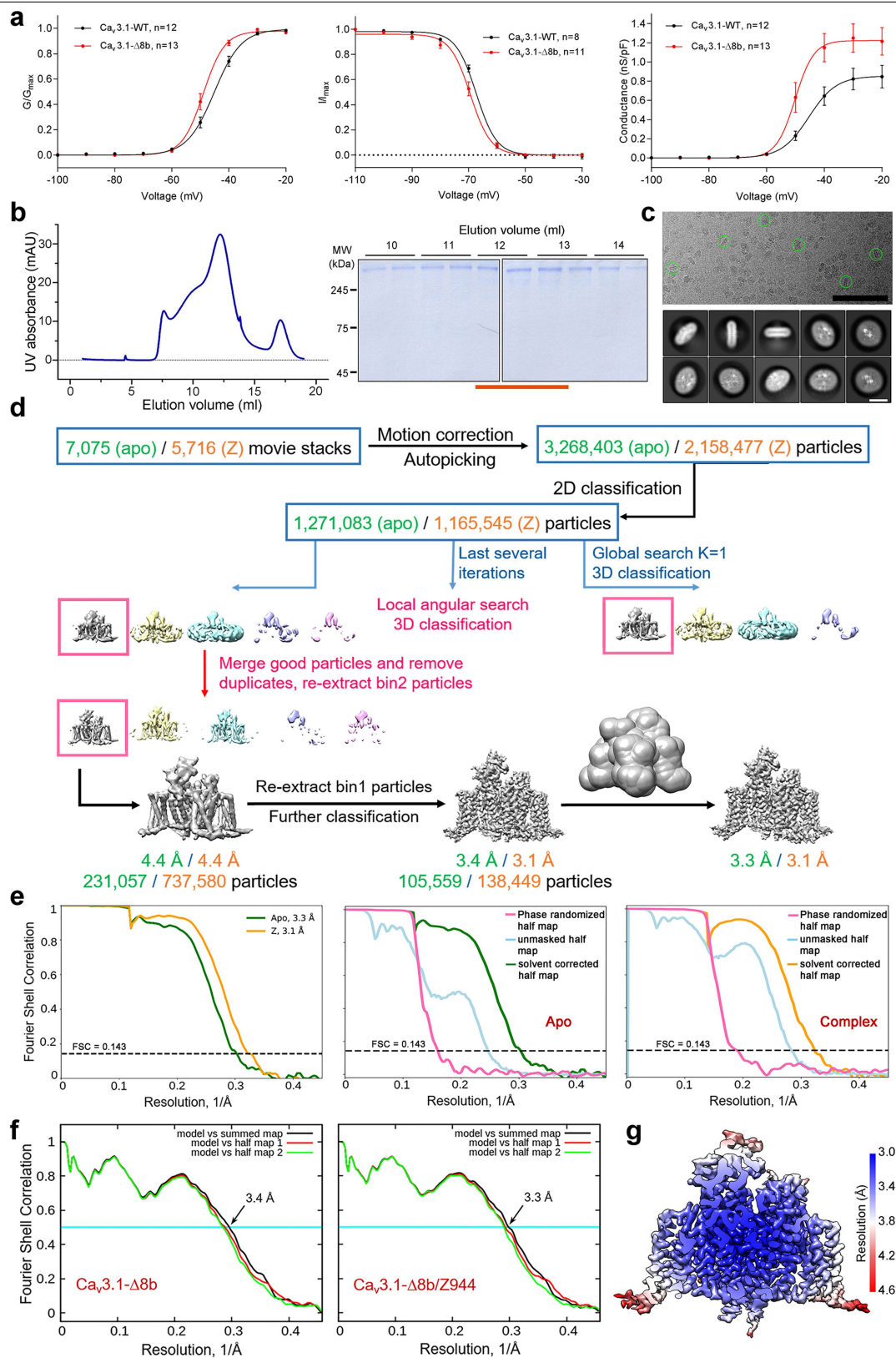
Extracellular view

Extended Data Fig. 1 | See next page for caption.

Article

Extended Data Fig. 1 | Brief introduction to Ca_v channels. **a**, A brief overview of the classification, physiology and pharmacology of mammalian Ca_v channels. The evolutionary distance is calculated by Clustal W⁵⁵. The table is summarized from several reviews^{4,20,56,57}. HVA, high-voltage-activated; LVA, low-voltage-activated; E-C coupling, excitation–contraction coupling; E-T coupling, excitation–transcription coupling. **b**, Pairwise comparison of sequence similarity and identity of full-length human Ca_v channels. The sequence alignment is provided as Supplementary Fig. 1. **c**, Topological structure of the Ca_v channels. The panel is adapted from our previous publication with some modifications²³. For Ca_v3.1-Δ8b, residues 509–642, shown as dashed lines on the I–II linker, were deleted. No human splice variant

corresponding to the mouse Ca_v3.1-Δ8b has been identified. In fact, the exon–intron boundaries do not support existence of such a variant in humans. Nevertheless, we name this construct Ca_v3.1-Δ8b to acknowledge the source where this construct was generated. Five glycosylation sites are observed on the extracellular loops, including Asn246/322/1428/1425/1675 (Fig. 1a). Glycosylation of the counterparts of Asn246 and Asn1425 has been reported in Ca_v3.2^{58,59}, and the glycosylation might modulate channel expression and activity^{60–62}. **d**, The typical domain-swapped architecture of most voltage-gated ion channels⁶³. Shown here is an extracellular view in which the voltage-sensing domains are shown as round rectangles.



Extended Data Fig. 2 | See next page for caption.

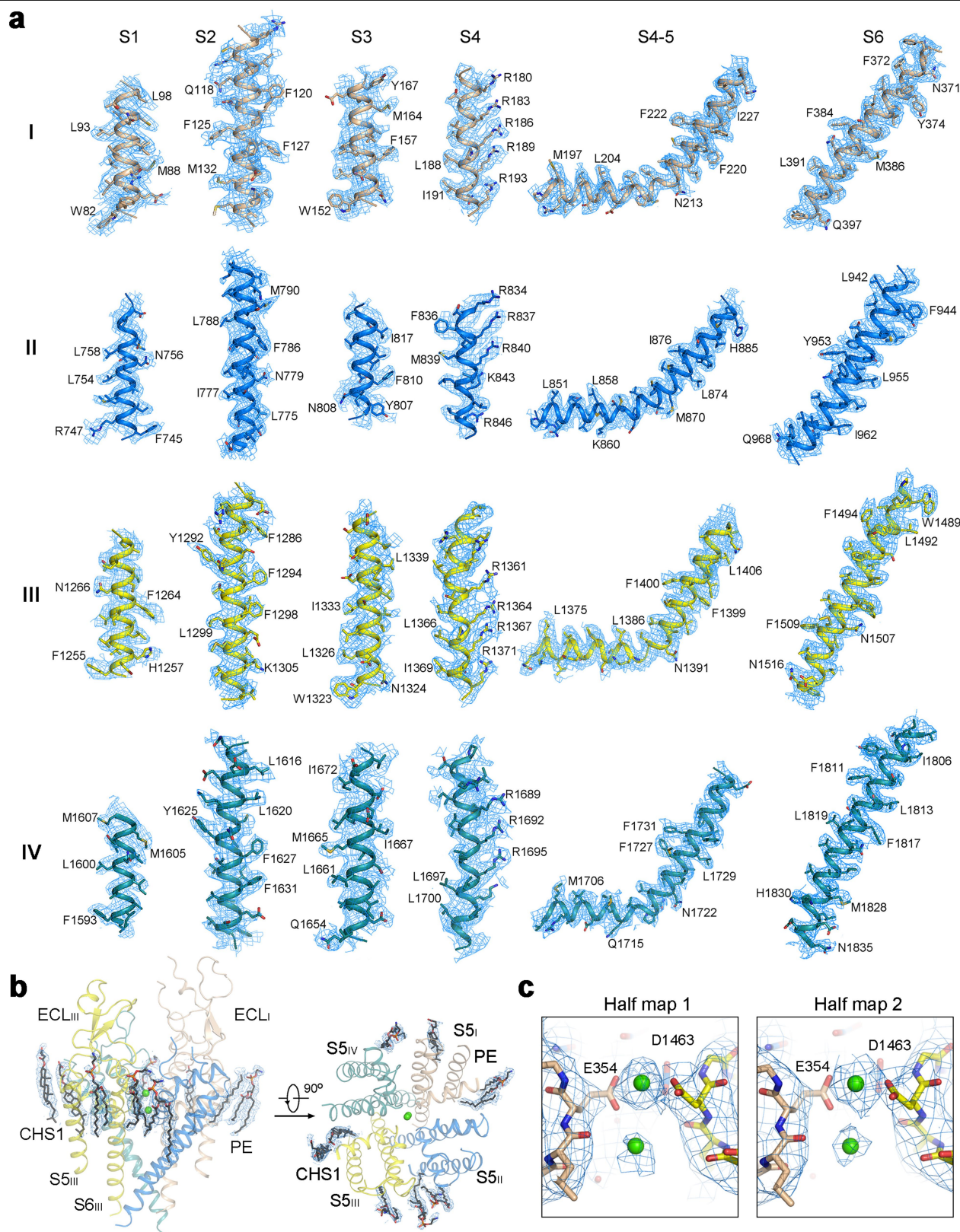
Extended Data Fig. 2 | Cryo-EM analysis of the human Ca_v3.1-Δ8b alone and in complex with Z944. **a**, Whole-cell patch clamp measurements of the full-length human Ca_v3.1 and Ca_v3.1-Δ8b. *n* values indicate the number of independent cells; mean ± s.e.m. **b**, Last-step purification of human Ca_v3.1-Δ8b.

Shown here is a representative size-exclusion chromatogram for proteins obtained from 30 l of HEK293F cells transfected with plasmids. The indicated peak fractions on the Coomassie-blue-stained SDS-PAGE (Supplementary Fig. 2) were pooled and concentrated for cryo-EM sample preparation.

c, Representative electron micrograph and 2D class averages. The green circles indicate representative particles in distinct orientations. The black and white scale bars in the top and bottom panels represent 100 nm and 10 nm, respectively.

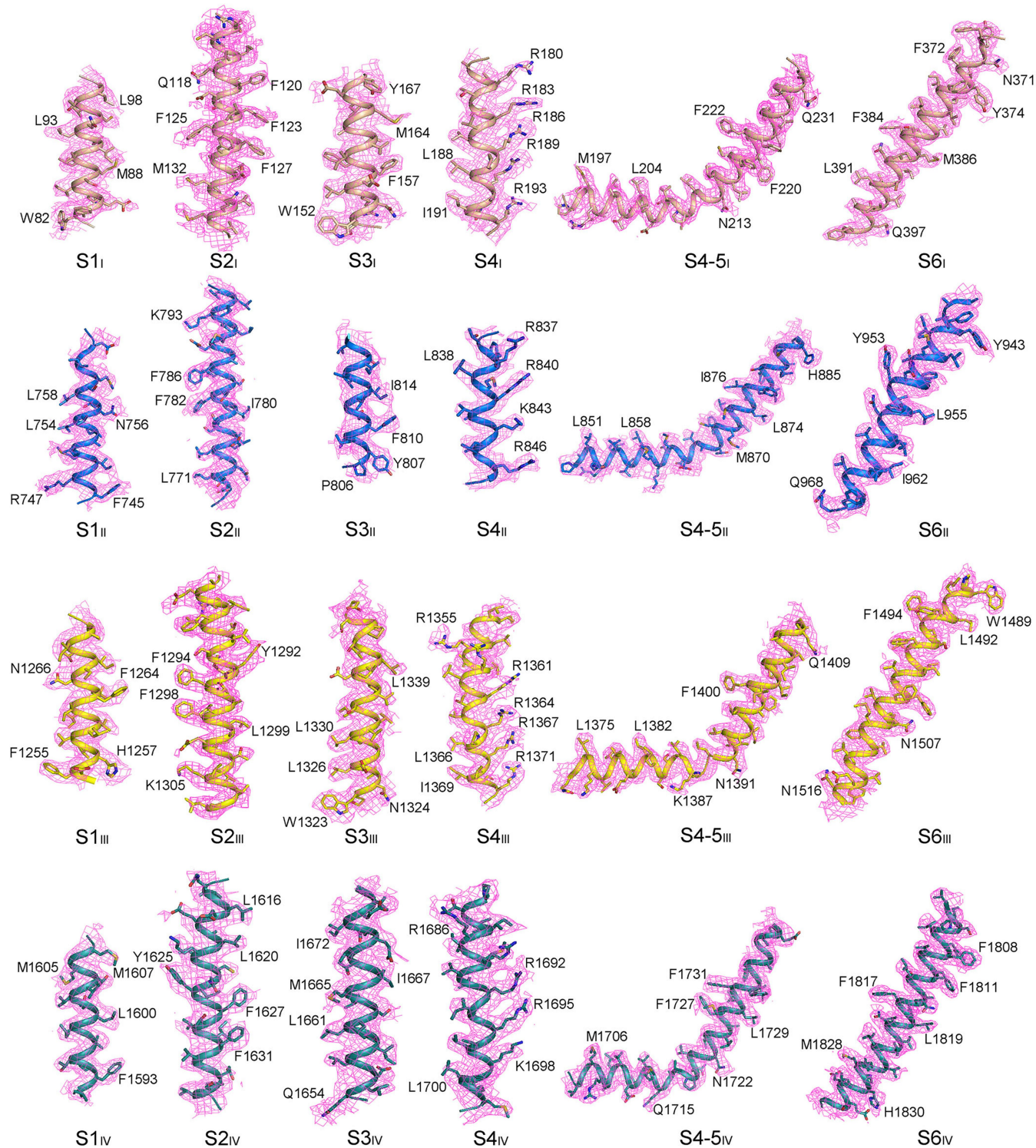
d, Flowchart for EM data processing. Details can be found in Methods. **e**, The gold-standard Fourier shell correlation (FSC) curves for the 3D reconstructions. The middle and right panels show FSC curves for phase-randomized half maps and unmasked half maps for the apo (middle) and complex (right) datasets. **f**, FSC curves of the refined model versus the overall

map that it was refined against (black); of the model refined in the first of the two independent maps used for the gold-standard FSC versus that same map (red); and of the model refined in the first of the two independent maps versus the second independent map (green). The small difference between the red and green curves indicates that the refinement of the atomic coordinates did not suffer from overfitting. Before calculation of FSC against model-generated map, both half maps and the merged map were multiplied by a solvent mask that only includes the protein region. The merged map was brought to a threshold at which the micelle is invisible and all transmembrane helices are visible. Dust points were manually removed using the hide dust function in Chimera. Caution was taken not to mask out the densities for the bound ligand and lipids. The map was then extended by 2 pixels and supplied with a soft edge width of 12 pixels using `relion_mask_create`. **g**, Local-resolution map for the 3D EM reconstruction of Ca_v3.1-Δ8b in the presence of Z944. The map, calculated in RELION-3.1, was generated in Chimera⁴⁹.

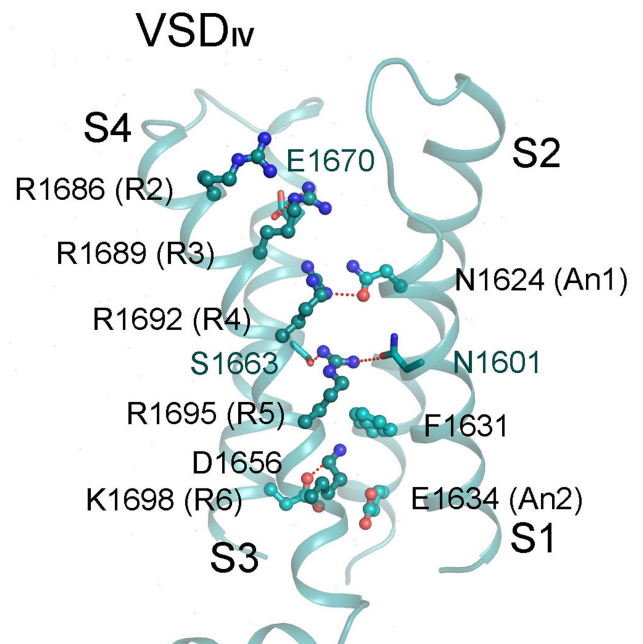
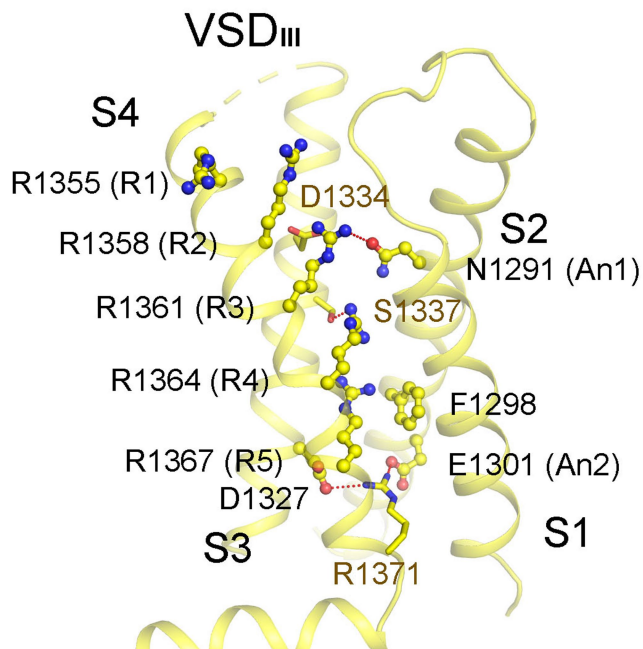
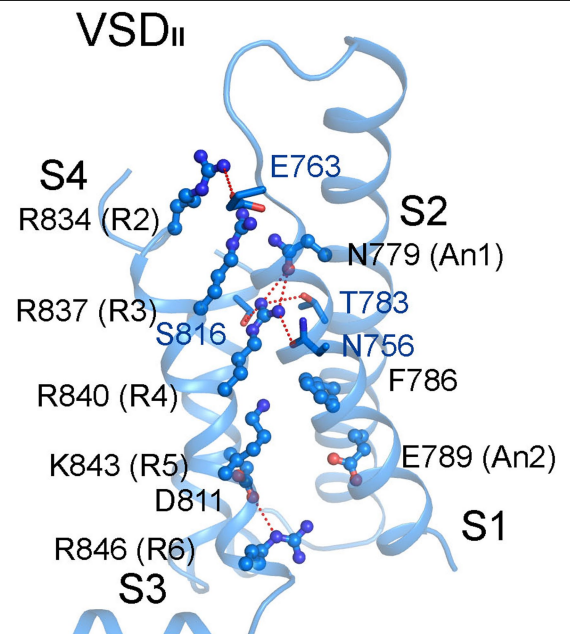
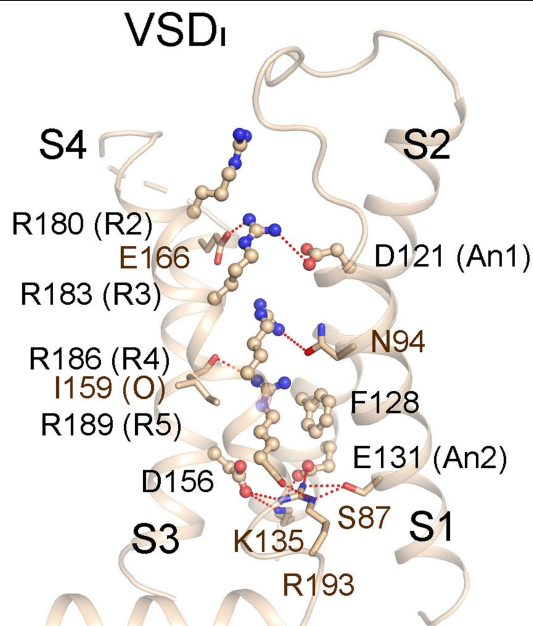


Extended Data Fig. 3 | EM maps for the transmembrane segments and lipids in Ca_v3.1-Δ8b. **a**, EM maps for the S1–S6 segments in each repeat, shown as blue mesh, are contoured at 4–5 σ . The maps were prepared in PyMol. **b**, Densities reminiscent of lipids and cholesteryl hemisuccinate (CHS) surrounding the

pore domain. The densities are contoured at 7 σ . For visual clarity, the ECL_I and ECL_{II} are omitted in the extracellular view. PE, phosphatidylethanolamine. **c**, EM densities for the two calcium ions and surrounding residues from two half maps. The densities are contoured at 4.5 σ .

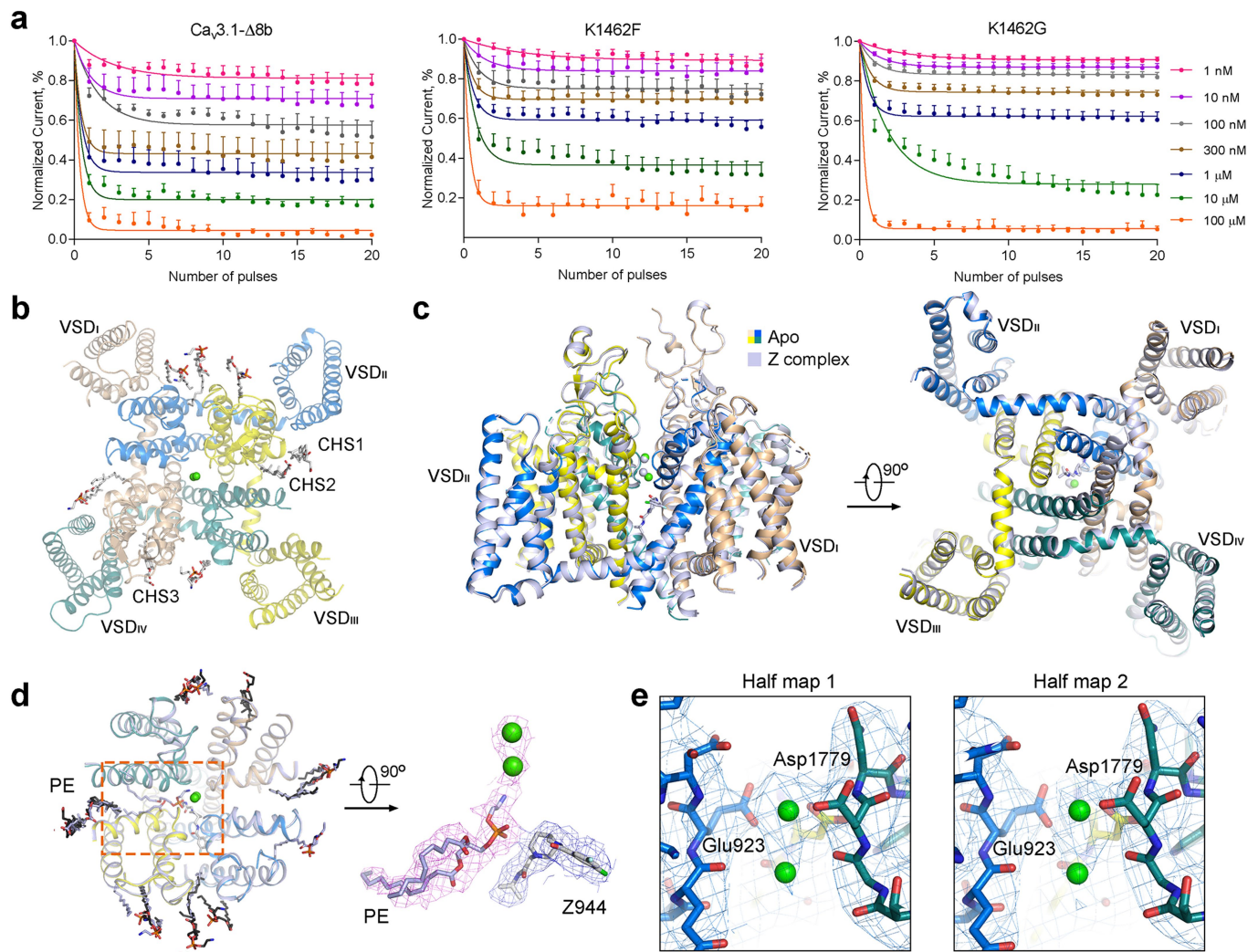


Extended Data Fig. 4 | EM maps for the transmembrane segments of the Z complex. EM maps for the S1–S6 segments in each repeat, shown as magenta mesh, are contoured at 4–5 σ .



Extended Data Fig. 5 | Depolarized ('up') conformations of the four VSDs. Structures of the four VSDs are presented in similar views. After purification in the absence of electric field with a lengthy duration, Na_v and Ca_v channels are expected to be trapped in the inactivated states that are featured with depolarized or 'up' VSDs and closed intracellular gate. The S4 segments are in

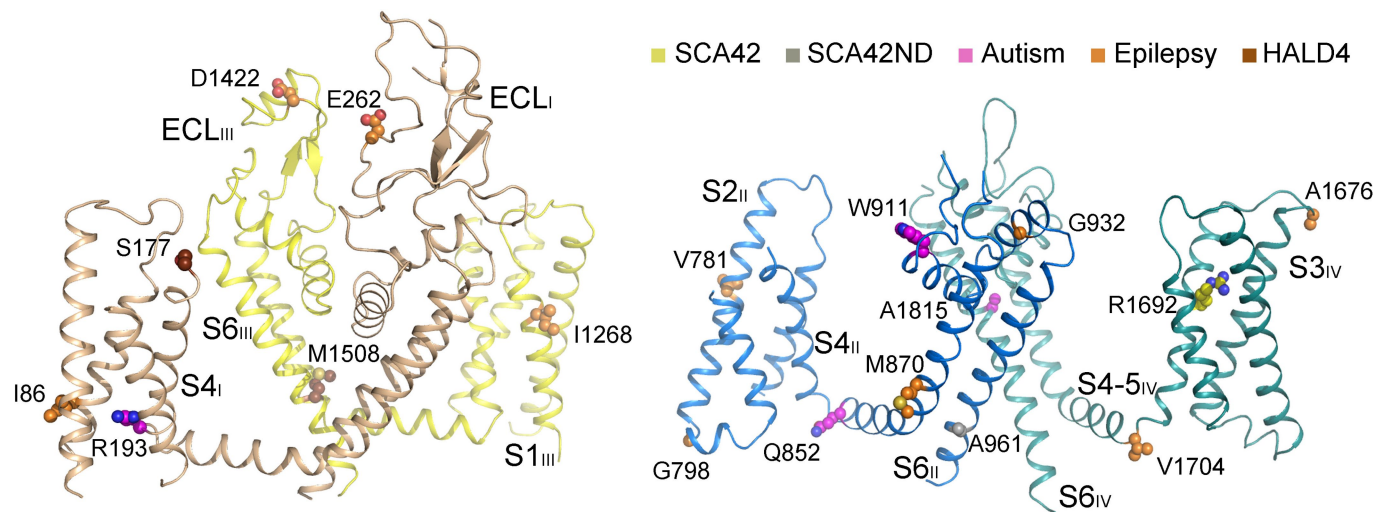
the 3₁₀ helix form. The gating charge residues and the conserved charge transfer centre⁶⁴ are shown as ball and sticks. Other polar residues that form potential hydrogen bonds are represented by red dashed lines, with the gating charge residues shown as sticks. The two conserved polar or acidic residues on S2 that facilitate charge transfer, designated An1 and An2, are also labelled.



Extended Data Fig. 6 | Local structural shifts of $\text{Ca}_v3.1-\Delta 8b$ upon Z944 binding.

a, Lys1462, which is conserved in T-type channels only, is important for Z944 inhibition. State-dependent blockade by Z944 at indicated concentrations in cells expressing $\text{Ca}_v3.1-\Delta 8b$ (left), $\text{Ca}_v3.1-\Delta 8b$ (K1462F) (middle) and $\text{Ca}_v3.1-\Delta 8b$ (K1462G) (right) are tested. n values indicate the number of independent cells; mean \pm s.e.m. The sample sizes (n) tested from low to high concentrations are: $n = 4, 5, 5, 3, 8, 6, 3$ for $\text{Ca}_v3.1-\Delta 8b$; $n = 3, 4, 8, 8, 8, 3, 3$ for $\text{Ca}_v3.1-\Delta 8b$ (K1462F); and $n = 8, 8, 10, 10, 8, 6, 3$ for $\text{Ca}_v3.1-\Delta 8b$ (K1462G). **b**, Several lipid and CHS molecules are resolved in the structure of $\text{Ca}_v3.1-\Delta 8b$. Shown here is an extracellular view. The lipids, the precise identities of which remain unclear, are shown as sticks. Phosphatidylethanolamine (PE) molecules were tentatively modelled into these densities. Three densities are reminiscent of cholesteryl hemisuccinate (CHS1–CHS3), although they may also belong to the detergent glyco-diosgenin (GDN). **c**, Structures of $\text{Ca}_v3.1-\Delta 8b$ alone and in complex with Z944 can be superimposed, with a r.m.s.d. of 0.45 Å over 851 Cα atoms. Two perpendicular views of the superimposed structures are shown. $\text{Ca}_v3.1-\Delta 8b$ alone is coloured by domain and the complex is coloured light blue. **d**, Change of lipid distribution in the pore domain in the presence of Z944. Left: an extracellular view of the superimposed pore domain of $\text{Ca}_v3.1-\Delta 8b$ with or without Z944. The bound lipids, shown as thin sticks, are coloured dark grey for

the apo structure and light blue for the complex. Z944 is shown as silver sticks. An extra lipid molecule, highlighted with a red rectangle, was resolved in the pore domain of the complex. Right: the densities for Z944 and the nearby transverse lipid are contoured at 4.5σ. It is noted that the densities that were tentatively assigned with two Ca^{2+} ions are contiguous with that for the transverse lipid. Although we cannot entirely exclude the possibility that the densities in the selectivity filter (SF) may belong to a lipid, they are more likely to be bound ions because: (1) If the density belongs to the head group of a lipid, the SF is too narrow to accommodate any known positively charged linear head group with the length corresponding to the density; if the density belongs to a tail, then the hydrophobic property is incompatible with the polar environment within the SF. (2) Lipid-like densities have been observed traversing the pore domain in nearly all structures of Na_v and Ca_v channels with fenestrations. In these channels, a highly conserved inner site constituted by backbone C=O groups has been demonstrated to coordinate Na^+ or Ca^{2+} by X-ray crystallographic and molecular dynamics simulation analyses. Taken together, two Ca^{2+} ions, instead of a lipid moiety, were tentatively assigned to the density in the SF. **e**, Half-map densities for the SF from two diagonal repeats, contoured at 4.5σ.



Extended Data Fig. 7 | Structural mapping of disease mutations identified in Ca,3.1 and Ca,3.2. Please refer to Extended Data Table 1 for details. Side views of the diagonal repeats are shown. SCA42, spinocerebellar ataxia 42; SCA42ND, spinocerebellar ataxia 42, early-onset, severe, with neurodevelopmental deficits; HALD4, hyperaldosteronism, familial, 4.

Extended Data Table 1 | Structural mapping of disease-related mutations identified in human T-type VGCC

Ca _v	Mutations	Disease	Structure	References	In Structure
Ca _v 3.1	R1692H	SCA42	S4 _{IV}	^{65,66}	R1692
	A961T	SCA42ND	S6 _{II}	⁶⁷	A961
	M1508V	SCA42ND	S6 _{III}	⁶⁷	M1508
Ca _v 3.2	R212C	Autism	S4 _I	⁶⁸	R193
	R902W	Autism	S4-5 _{II}	⁶⁸	Q852
	W962C	Autism	P1 _{II}	⁶⁸	W911
	A1847V	Autism	S6 _{IV}	⁶⁸	A1815
	V105G	Epilepsy	S1 _I	⁶⁹	I86
	F161L	Epilepsy	S2-3 _I	⁷⁰	--
	Q163H	Epilepsy	S2-3 _I	⁶⁹	--
	E282K	Epilepsy	Extracellular I	⁷⁰	E262
	A332T	Epilepsy	Extracellular I	⁶⁹	--
	C456S	Epilepsy	I-II linker	⁷⁰	--
	A480T	Epilepsy	I-II linker	⁷¹	--
	G499S	Epilepsy	I-II linker	⁷⁰	--
	P618L	Epilepsy	I-II linker	⁷¹	--
	P648L	Epilepsy	I-II linker	^{70,72}	--
	R744Q	Epilepsy	I-II linker	⁷⁰	--
	A748V	Epilepsy	I-II linker	⁷⁰	--
	G755D	Epilepsy	I-II linker	⁷¹	--
	G773D	Epilepsy	I-II linker	⁷⁰	--
	G784S	Epilepsy	I-II linker	⁷⁰	--
	V831M	Epilepsy	S2 _{II}	⁷⁰	V781
	G848S	Epilepsy	S2 _{II}	⁷⁰	G798
	A876T	Epilepsy	S3-4 _{II}	⁶⁹	--
	T920M	Epilepsy	S5 _I	⁶⁹	M870
	G983S	Epilepsy	P2 _{II}	⁶⁹	G932
	A1059S	Epilepsy	II-III linker	⁶⁹	--
	E1170K	Epilepsy	II-III linker	⁶⁹	--
	Q1264H	Epilepsy	II-III linker	⁶⁹	--
	V1309I	Epilepsy	S1 _{III}	⁶⁹	I1268
	D1463N	Epilepsy	Extracellular III	⁷⁰	D1422
	T1606M	Epilepsy	S0 _{IV}	⁶⁹	--
	A1705T	Epilepsy	S3-4 _{IV}	⁶⁹	A1676
	T1733A	Epilepsy	S4-5 _{IV}	⁶⁹	V1704
	R1892H	Epilepsy	CTD	⁶⁹	--
	A1966V	Epilepsy	CTD	⁶⁹	--
	R2005C	Epilepsy	CTD	⁶⁹	--
	A2140T	Epilepsy	CTD	⁶⁹	--
	A2170T	Epilepsy	CTD	⁶⁹	--
	M2312V	Epilepsy	CTD	⁶⁹	--
	S196L	HALD4	S4 _I	⁷³	S177
	M1549I/V	HALD4	S6 _{III}	^{73,74}	M1508
	P2083L	HALD4	CTD	⁷³	--
Ca _v 3.3	Mutation frequency not determined				

The mutations are summarized from <https://www.uniprot.org/uniprot/O43497>, <https://www.uniprot.org/uniprot/O95180> and references⁶⁵⁻⁷⁴. UNIPROT number: Ca_v3.1, O43497-9; Ca_v3.2, O95180-1. SCA42, spinocerebellar ataxia 42; SCA42ND, spinocerebellar ataxia 42, early-onset, severe, with neurodevelopmental deficits; HALD4, hyperaldosteronism, familial, 4. The mutations that can be structurally mapped are shaded by light colours coded for disease types.

Extended Data Table 2 | Activation, steady-state inactivation and conductance parameters of Ca_v3.1 variants transiently expressed in HEK293T cells

	Parameters	Ca _v 3.1-WT	Ca _v 3.1-Δ8b	K1462G	K1462F
Activation	V _{1/2} (mV)	-45.10 ± 0.44	-48.90 ± 0.39 ^{***}	-36.40 ± 0.54 ^{***}	-41.80 ± 0.60 ^{***}
	P	/	0.0000000017	<0.000000000000001	<0.000000000000001
	slope	4.74 ± 0.34	3.71 ± 0.42	6.48 ± 0.45 ^{***}	5.77 ± 0.52 ^{**}
	P	/	/	0.000006	0.0017
	n	12	13	18	13
Steady-state inactivation	V _{1/2} (mV)	-67.40 ± 0.23	-69.47 ± 0.43 ^{***}	-66.20 ± 0.46 ^{***}	-67.45 ± 0.23 ^{***}
	P	/	0.0187	0.000001	0.000017
	slope	-3.08 ± 0.20	-4.32 ± 0.45	-4.81 ± 0.38	-3.73 ± 0.20
	τ _{inac}	/	18.49 ± 0.75	10.41 ± 0.41 ^{***}	10.45 ± 0.74 ^{***}
	P	/	/	0.00000000043	0.0000000002
	n	8	11	11	11
Conductance	G _{top} (nS / pF)	0.85 ± 0.06	1.23 ± 0.07 ^{**}	/	/
	P	/	0.0058	/	/
	n	12	13	/	/

In the comparison of Ca_v3.1-WT with Ca_v3.1-Δ8b, ** (red) $P < 0.01$ versus WT, *** (red) $P < 0.001$ versus WT. In the comparison of Ca_v3.1-Δ8b with Ca_v3.1-Δ8b (K1462G) and Ca_v3.1-Δ8b (K1462F), ** $P < 0.01$ versus Ca_v3.1-Δ8b, *** $P < 0.001$ versus Ca_v3.1-Δ8b. Each data point represents mean ± s.e.m. (standard deviation of mean). n indicates the number of independent cells. The extra sum-of-squares F test was used to compare the V_{1/2} of activation and inactivation and G_{top} of conductance fits. τ_{inac} values of Ca_v3.1-Δ8b, Ca_v3.1-Δ8b (K1462G) and Ca_v3.1-Δ8b (K1462F) were compared by one-way ANOVA and Tukey's test. The data were analysed using Prism 8.2.1 (GraphPad Software).

Extended Data Table 3 | Statistics for data collection and structural refinement

	Cav3.1-Δ8b (EMD-0791) (PDB 6KZO)	Cav3.1-Δ8b/Z944 (EMD-0792) (PDB 6KZP)
Data collection and processing		
Magnification	× 105,000	× 105,000
Voltage (kV)	300	300
Electron exposure (e-/Å ²)	48	48
Defocus range (μm)	-1.3~-1.8	-1.3~-1.8
Pixel size (Å)	1.091	1.091
Symmetry imposed	C1	C1
Initial particle images (no.)	3,268,403	2,158,477
Final particle images (no.)	105,559	138,449
Map resolution (Å)	3.3	3.1
FSC threshold	0.143	
Map resolution range (Å)	3.2-50	2.9-50
Refinement		
Initial model used (PDB code)	6J8E	6J8E
Model resolution (Å)	3.4	3.3
FSC threshold	0.143	
Model resolution range (Å)	3.2-50	2.9-50
Map sharpening <i>B</i> factor (Å ²)	-120.0	-120.0
Model composition		
Non-hydrogen atoms	8248	8301
Protein residues	984	984
Ligands	9	12
<i>B</i> factors (Å ²)		
Protein	67.93	43.19
Ligand	62.22	34.49
R.m.s. deviations		
Bond lengths (Å)	0.005	0.006
Bond angles (°)	0.997	1.043
Validation		
MolProbity score	1.52	1.35
Clashscore	4.57	2.39
Poor rotamers (%)	0.35	0.46
Ramachandran plot		
Favored (%)	95.82	95.32
Allowed (%)	4.08	4.47
Disallowed (%)	0.10	0.21

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|--|
| n/a | Confirmed |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of all covariates tested |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection AutoEMation; Patchmaster 2.90.4;

Data analysis MotionCorr; MotionCor2 1.1.0; GCTF 1.06; RELION 2.1; RELION-3.0-beta; Chimera 1.13; Coot 0.8.6.1; Phenix 1.13; Pymol 1.8.6.0; Fitmaster 2.90.4; Prism 8.2.1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Both structures are cryo-EM structures in our manuscript. The structures and maps will be submitted to PDB and EMDB soon. The data availability statement has been provided in the manuscript as follow:

The atomic coordinates and EM maps for Cav3.1 alone and in complex with Z944 have been deposited in the Protein Data Bank (<http://www.rcsb.org>) with the accession codes 6KZO and 6KZP, and EMDB (<https://www.ebi.ac.uk/pdbe/emdb/>) with the codes EMD-0791 and EMD-0792, respectively.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were chosen to give s.e.m. values of less than 10% of peak values based on prior experimental experience.
Data exclusions	For the ion current measurement, data over 10 nA or less than 500 pA were excluded to ensure the currents be properly voltage-clamped and avoid potential endogenous channel contamination.
Replication	For the ion current measurement, each experiment was replicated for at least 3 times. All attempts at replication were successful.
Randomization	The GFP positive cells were randomly selected for whole-cell patch.
Blinding	All the constructs were recorded and analyzed blindly to avoid bias.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	HEK293F (invitrogen); HEK293T (ATCC)
Authentication	No further authentication was performed for commercially available cell lines.
Mycoplasma contamination	The cell lines were not tested for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used.



NICK GRAHAM

Christina Hicks interviews a fisher in Kenya during her PhD.

WHAT I LEARNT FROM BEING A LEAD AUTHOR

Scientists reveal key lessons from the publishing process. **By Chris Woolston**

With a new year approaching, researchers everywhere are taking stock of their work and their future. Even for those who had successes this year, 2020 holds uncertainty as well as promise. We asked scientists who were first-time lead authors on a paper published in *Nature* or a *Nature* journal in 2019 to talk about their careers and lessons they have learnt.

CHRISTINA HICKS MAKE YOUR STORY A COMPELLING ONE

I had the idea for the paper four years ago when I was on maternity leave. I didn't know where I wanted to go with my career. I wanted to do something that would have a real-world impact. As I thought about it, I realized that I

could link fisheries to food insecurity. You have to be passionate about your idea to get past the stumbling blocks. Passion gives you stamina.

It's important to be with people you like and trust. Two of my co-authors on the *Nature* paper are my best friends from my PhD programme at James Cook University in Townsville, Australia (*Nature* **574**, 95–98; 2019). My husband and one of his best friends are also co-authors.

It's not an accident. I spend so much time

Work / Careers

thinking about work that my colleagues are also my friends.

Collaboration changes the way you think about your work. One of my co-authors is a nutritionist who specializes in child and maternal health. Nutritionists use a level of precision in their work that you don't usually see in fisheries science or ecology. When he raised an issue, I would think, 'That's ridiculous; why can't he just move on?'

But six months down the line, I realize how significant it really was.

We considered breaking this research into three papers. In the end, we decided to combine the three parts into one paper, and I'm glad we did. If you want to hit a big journal, you need to tell a compelling story, and that often means putting multiple pieces of a puzzle together.

I also wanted the work to reach the broadest possible audience. If your work is published in a high-impact journal, it's a lot easier to talk to policy- and decision-makers in government.

What happened next?

The paper hasn't been out very long, but I can tell it's resonating through diverse disciplines. I've heard from chemists, fishery biologists and people who study chemical stoichiometry. That's exactly what I hoped would happen.

What's your career goal for 2020?

I won a grant from the European Research Council last year, so actively hiring postdocs and PhD students. I want to build a group that's ambitious and productive but also recognizes the importance of work-life balance. For my research, I want to learn why so many people in East Africa and elsewhere are unable to access the nutrients from fisheries.

What is your biggest lesson?

I need to focus on myself for me. It's easy to become really busy. I run up and down the mountains in the UK Lake District for fun. My brain gets cluttered and overworked if I don't take a break. Running helps me to slow down.

Christina Hicks is an environmental social scientist at Lancaster University, UK.

JOHAN VANDEN HOOGEN KNOW THE VALUE OF OUTSOURCING

I was pretty cynical about science and my career when I was doing my PhD at Wageningen University in the Netherlands. I was convinced that I would never even be a co-author on a paper in a top-tier journal. Until the third year of my PhD programme, which I completed last year, I didn't have a single manuscript that was ready to submit for publication. I finally



JORIS SCHAAP

Johan van den Hoogen takes a sample of soil.

published a paper at the end of my fourth year, and that made a big difference to my career and my confidence. You have to show yourself that you can finish something.

I came to an important realization that year: I don't really care too much about having my own group. Since I've been at the Swiss Federal Institute of Technology (ETH) Zurich, people have been asking when I'm going to become a professor. But it's funny. I have a friend who works at a pharmaceutical company, and I never ask him when he's going to become the chief executive or the head of his department.

My title is senior scientist, but I'm a jack of all trades. I help other people with data collection and cleaning, and with writing codes and papers, but I still have enough time for my own research. I see that some principal investigators spend only 10–20% of their time doing research. I wouldn't be happy with that.

This year, I managed to be first author on a *Nature* paper (*Nature* 572, 194–198; 2019). The nematode project that inspired it had been running for a year before I joined. I knew little about nematodes at the time. In this particular case, knowing a little bit less about the topic might have actually helped – we were creating a global map, so I needed a big-picture view.

If you know a lot about what happens in your backyard, you can get lost in the details. I didn't even know the details.

What happened next?

Getting that paper published didn't set off the massive celebration that you might expect. That's partly because we had a lot of small celebrations every step of the way. It's great to have a first-author paper in *Nature*. It makes people aware of my work. But as far as my career goes, it wouldn't have made much difference if it had been published in a lower-impact journal. I'm not chasing publications.

What is your career goal for 2020?

I'm perfectly happy where I am at ETH Zurich. I can do the work that I want to do without having to stress about getting funding or moving to another laboratory or another country in a year and a half. Maybe I'll do something completely different in 5 or 10 or 15 years, but I don't have to worry about that now.

What is your biggest lesson?

Working on that nematode paper helped me to appreciate the value of outsourcing – I didn't develop the models in it. You should let other

people do the things that they're good at. I understand the models in the paper, but it would have taken me a year and half to create them on my own.

My biggest realization is that you don't need to move up the academic career ladder to have a satisfying career in science. The moment I stopped worrying about advancing in academia marked a change for me.

Johan van den Hoogen is a soil ecologist at ETH Zurich, Switzerland.

STEPHANIE ELLIS

ACADEMIA OR BUST

I did my PhD on fruit flies in a small, relatively new laboratory at the University of British Columbia in Vancouver, Canada. Now I'm in a large, high-powered lab where I have a lot of flexibility to do what I want. To really stand out, I needed to shift my perspective. I realized that technologies recently developed in the lab would allow me to study cell competition – a sort of survival of the fittest – in mouse skin cells, something that no one else in the lab was working on.

My supervisor, Elaine Fuchs, has been extremely supportive, but I had to get her excited about the project and convince her that this was an important problem. Then I had to develop fresh angles to the story to get other people in the lab excited. I have colleagues from a lot of different backgrounds, so I needed to explain it in a way that resonated with everyone. If you are having trouble getting people excited about your work, you need to change your thinking.

The cell-competition paper had three authors when we first submitted it, and it ended up with six (*Nature* **569**, 497–502; 2019). One reviewer suggested that we do a single-cell

RNA experiment that I really didn't want to do. I had to start a new collaboration, and after four or five months of back and forth, we came up with another idea. I never would have done that experiment on my own. It pushed me out of my comfort zone.

What happened next?

When you publish in a high-impact journal, you have to be prepared for the aftermath. A lot of people are reading and talking about that paper. Some commenters on Twitter tried to minimize it by pointing out that cell competition is already well documented in fruit flies. There will always be naysayers, but I'm proud of the work.

What is your career goal for 2020?

I'm at the end of a six-year postdoctoral position and am now applying for faculty jobs. I've had a good application season so far, and it's because of the paper. My only goal is to have my own lab. Since I did my first fruit-fly experiment in graduate school, I've been hooked. For me, it's academia or bust.

What is your biggest lesson?

You have to have people around you who can point out the weaknesses in anything you do.

Stephanie Ellis is a cell biologist at the Howard Hughes Medical Institute, Rockefeller University, New York City, New York.

OSCAR SERRANO

CHOOSE COLLABORATORS CAREFULLY

My paper had 45 authors, so it was a very collaborative effort (*Nature Commun.* **10**, 4313; 2019). We needed to collect all of the available data on blue carbon – carbon that is sequestered by coastal communities of mangroves,

sea grasses and tidal marshes – from around Australia. Scientists aren't always willing to share unpublished data, but in this case there wasn't much hesitation. The other researchers were aware that blue carbon is a hot topic and saw the value of the project. We offered co-authorship as an incentive to everyone who contributed data.

We had a clear goal for the paper, but there were still a lot of opinions, comments and strong wills. After the first round of suggestions, I encouraged co-authors to focus on the big things because it's not possible to incorporate every wish and every little change from every author. But not all of them followed that advice. It was funny to see how many co-authors would try to improve the manuscript even though it was already in good shape. It was an exhausting process.

But in the end, improving it is what it's all about. When someone really got into the paper, I knew that I could collaborate with that person in the future.

What happened next?

Because of that paper, companies are contacting me for more information and advice about investing in carbon credits. I also got some media attention on television and the radio and in the newspapers. I was able to talk about the importance of these ecosystems for carbon sequestration and climate change. When speaking to the media, you learn to get straight to the point and avoid a lot of jargon. You also have to be friendly, like you're talking to your neighbour.

What is your career goal for 2020?

I'm still trying to secure a position for next year. Among other things, I'm looking for opportunities to return to my home country – Spain – where I still have friends and family. I want to remain in academia but continue to interact with industry to help preserve these ecosystems. It's a difficult career. I'm 38 and I still don't have a stable, long-term position.

What is your biggest lesson?

Some relationships are more mutually beneficial than others. You can collaborate for years with someone and then realize it's been a one-way street. You're sharing ideas and resources with them but getting little in return. But other people really do give back as much as they get and really help you grow your career. I want to be the person that people want to collaborate with because it's reciprocal.

Oscar Serrano is a marine ecologist at Edith Cowan University in Joondalup, Australia.

Interviews by Chris Woolston.

These interviews have been edited for length and clarity.



Oscar Serrano works on carbon sequestration.



Where I work Jodi Rowley

As an amphibian biologist, I'm particularly obsessed with frogs, which are important in ecosystems because they eat insects and are themselves eaten by birds and reptiles. But frogs are vulnerable: 40% of amphibian species worldwide are at risk of extinction. We've already lost at least four frog species in Australia. By going into the field to assess frog populations and identify new species, I collect data that help them to get the attention they need for conservation planning.

This photo was taken last September, when I was teaching at a field school in the Macquarie Marshes in New South Wales. Given Australia's ongoing drought, the name 'Marshes' seemed like a joke: it was more of a desert. But during a flood, tons of frogs emerge from the ground.

This trip was still really exciting because, for the first time, we got to see how the frogs cope with drought. We saw mostly barking marsh frogs (*Limnodynastes fletcheri*), and discovered that they hide in cracks in the ground of dry riverbeds, hoping that any

moisture there will last until the next rains or floods.

This trip made me think that climate change might be affecting these animals. It inspired me, in future research, to try to understand frogs' strategies for surviving dry conditions, how many live under the ground and how long they can survive without water.

It's so important to get out of the office and actually see the animals. That said, field expeditions have been some of the hardest times of my life. On this last trip, I got eight ticks. In Vietnam, I've been in rains and floods, holding a tarp over my hammock to keep it dry during storms. On a trip to the Solomon Islands, spending up to nine and a half hours a day climbing mountains, I thought I'd drop dead with tiredness.

It's tough – but fieldwork helps to keep my passion for frogs and conservation alive.

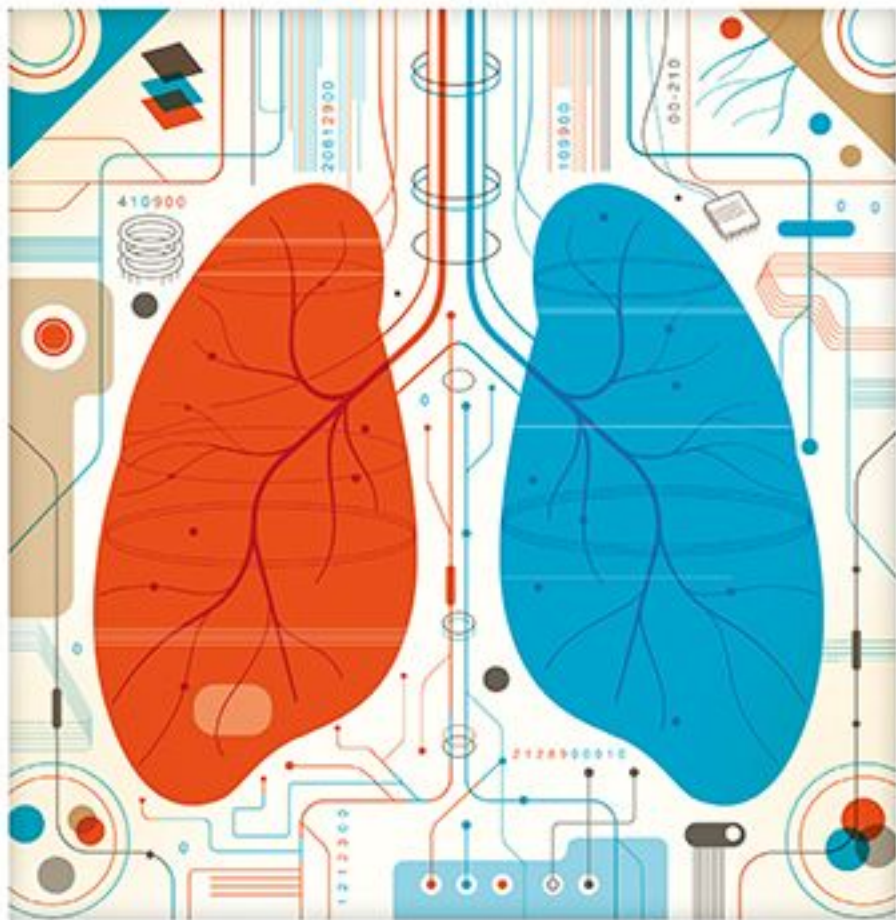
Jodi Rowley is curator for amphibian and reptile conservation biology at the Australian Museum and the University of New South Wales in Sydney. **Interview by Amber Dance.**

Photographed for *Nature*
by Tharindu Jay

Innovations^{by}

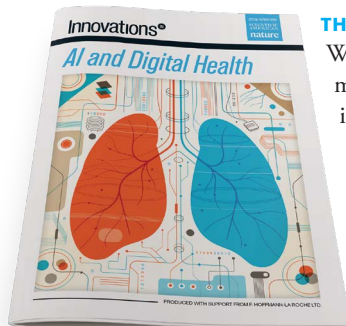
SPECIAL REPORT FROM
SCIENTIFIC
AMERICAN
nature

AI and Digital Health



PRODUCED WITH SUPPORT FROM F. HOFFMANN-LA ROCHE LTD.

How Artificial Intelligence Will Change Medicine



THE BIOMEDICAL WORLD IS AWASH IN DATA.

We have terabytes of genomic information from mouse to human, troves of health metrics from clinical trials, and reams of so-called real-world data from insurance companies and pharmacies. Using powerful computers, scientists have scrutinized this bounty with some fine results, but it has become clear that we can learn much more with an assist from artificial intelligence. Over the next decade deep-learning neural networks will likely transform how we look for patterns in data

and how research is conducted and applied to human health. This special report explores the promise of this nascent revolution.

Right now the biggest bets are being placed in the realm of drug discovery (page S49). And for good reason. The average cost of bringing a new drug to market nearly doubled between 2003 and 2013 to \$2.6 billion, and because nine out of 10 fail in the final two phases of clinical trials, most of the money goes to waste. Every large pharma company is working with at least one AI-focused start-up to see if it can raise the return on investment. Machine-learning algorithms can sift through millions of compounds, narrowing the options for a particular drug target. Perhaps more exciting, AI systems—unconstrained by prevailing theories and biases—can identify entirely new targets by spotting subtle differences at the level of tissues, cells, genes or proteins between, say, a healthy brain and one marked by Parkinson's—differences that might elude or even mystify a human scientist.

That same sharp-eyed ability is also being deployed to interpret medical scans (page S54). Some systems can already detect early signs of cancer that might be missed by a radiologist or see things that are simply beyond human capacity—such as assessing cardiovascular risk from a retinal scan. The Food and Drug Administration is approving imaging algorithms at a rapid clip. Other AI applications lie a bit further down the road. Will the inefficiencies of today's electronic health records (EHRs) be addressed by smart systems that prevent prescribing errors and provide early warnings of disease? Some of the world's biggest tech giants are working on it (page S59).

Despite fears that machines will displace humans, most experts believe artificial and human intelligence will work synergistically. The bigger concern is a shortage of people with both biomedical knowledge and algorithm-building proficiency (page S64). If this human problem can be resolved, the key to creating successful AI applications may depend on the quality and quantity of what we feed their hungry maw. "We rely on three things," says the CEO of one deep-learning start-up. "Data, data and more data."

This report, published in *Scientific American* and *Nature*, is sponsored by F. Hoffmann-La Roche Ltd. It was produced independently by the editors of *Scientific American*, who take sole responsibility for the editorial content.

Claudia Wallis, Contributing Editor

S49 Hunting for New Drugs with AI

The pharmaceutical industry is in a drug-discovery slump. How much can AI help?
By David H. Freedman

S53 GRAPHIC: SPEEDING UP THE SEARCH FOR DRUGS

S54 Rise of Robot Radiologists

Deep-learning algorithms are peering into MRIs and x-rays with unmatched vision, but who is to blame when they make a mistake?
By Sara Reardon

S59 Can AI Fix Medical Records?

Digitized patient charts were supposed to revolutionize medical practice. Artificial intelligence could help unlock their potential.
By Cassandra Willyard

S62 Wiring Minds

Successfully applying AI to biomedicine requires innovators trained in contrasting cultures.
By Amit Kaushal and Russ B. Altman

EDITORIAL

ACTING EDITOR IN CHIEF
Curtis Brainard

CHIEF FEATURES EDITOR
Seth Fletcher

CONTRIBUTING EDITOR
Claudia Wallis

SENIOR EDITOR
Madhusree Mukerjee

SENIOR EDITOR
Jen Schwartz

SENIOR EDITOR
Kate Wong

CREATIVE DIRECTOR
Michael Mrak

SENIOR GRAPHICS EDITOR
Jen Christiansen

ASSOCIATE GRAPHICS EDITOR
Amanda Montañez

COPY DIRECTOR
Maria-Christina Keller

SENIOR COPY EDITORS
Daniel C. Schlenoff,
Aaron Shattuck,
Angélique Rondeau

MANAGING
PRODUCTION EDITOR
Richard Hunt

PREPRESS AND
QUALITY MANAGER
Silvia De Santis

PUBLISHER AND VP
Jeremy A. Abbate

CORPORATE PARTNERSHIPS
DIRECTOR,
NATURE RESEARCH
David Bagshaw



Hunting for New Drugs with AI

The pharmaceutical industry
is in a drug-discovery slump.
How much can AI help?

By David H. Freedman

THERE ARE MANY REASONS that promising drugs wash out during pharmaceutical development, and one of them is cytochrome P450. A set of enzymes mostly produced in the liver, CYP450, as it is commonly called, is involved in breaking down chemicals and preventing them from building up to dangerous levels in the bloodstream. Many experimental drugs, it turns out, inhibit the production of CYP450—a vexing side effect that can render such a drug toxic in humans.

Drug companies have long relied on conventional tools to try to predict whether a drug candidate will inhibit CYP450 in patients, such as by conducting chemical analyses in test tubes, looking at CYP450 interactions with better-understood drugs that have chemical similarities, and running tests on mice. But their predictions are wrong about a third of the time. In those cases, CYP450-related toxicity may come to light only during human trials, resulting in millions of dollars and years of effort going to waste. This costly inaccuracy can, at times, feel like “the bane of our existence,” says Saurabh Saha, senior vice president of research and development and translational medicine at Bristol-Myers Squibb.

Inefficiencies such as this one contribute to a larger problem: the \$1-trillion global pharmaceutical industry has been in a drug development and productivity slide for at least two decades. Pharmaceutical companies are spending more and more—the 10 largest ones now pay nearly \$80 billion a year—to come up with fewer and fewer successful drugs. Ten years ago every dollar invested in research and development saw a return of 10 cents; today it yields less than two cents. In part, that is because the drugs that are easiest to find and that safely and effectively treat common disorders have all been found; what is left is hunting for drugs that address problems with complex and elusive solutions and that would treat disorders affecting only tiny portions of the population—and thus could return far less in revenue.

Because finding new, successful drugs has become so much harder, the average cost of bringing one to market nearly doubled between 2003 and 2013 to \$2.6 billion, according to the Tufts Center for the Study of Drug Development. These same challenges have increased the lab-to-market time line to 12 years, with 90 percent of drugs washing out in one of the phases of human trials.

It's no wonder, then, that the industry is enthusiastic about artificial-intelligence tools for drug development. These tools do not work by having expert-developed analytical techniques programmed into them; rather users feed them sample problems (a molecule) and solutions (how the molecule ultimately behaves as a drug) so that the software can develop its own computational approaches for producing those same solutions.

Most AI-based drug-discovery applications take the form of a technique called machine learning, including a subset of the approach called deep learning. Most machine-learning programs can work with small data sets that are organized and labeled, whereas deep-learning programs can work with

raw, unstructured data and require much larger volumes. Thus, a machine-learning program might learn to recognize the different features of a cell after being shown tens of thousands of examples of photographs of cells in which the parts are already labeled. A deep-learning version can figure out those parts on its own from unlabeled cell images, but it might need to look at a million of them to do it.

Many scientists in the field think that AI will ultimately improve drug development in several ways: by identifying more promising drug candidates; by raising the “hit rate,” or the percentage of candidates that make it through clinical trials and gain regulatory approval; and by speeding up the overall process. A machine-learning program recently deployed by Bristol-Myers Squibb, for instance, was trained to find patterns in data that correlate with CYP450 inhibition. Saha says the program boosted the accuracy of its CYP450 predictions to 95 percent—a sixfold reduction in the failure rate compared with conventional methods. These results help researchers quickly screen out potentially toxic drugs and focus instead on candidates that have a stronger shot at making it all the way through multiple human trials to U.S. Food and Drug Administration approval. “Where AI can make a huge difference is having drugs that fail early on, before we make all that investment in them,” says Vipin Gopal, chief data and analytics officer at Eli Lilly.

Resources are now piling into the field. AI-based drug-discovery start-ups raised more than \$1 billion in funding in 2018, and as of last September, they were on track to raise \$1.5 billion in 2019. Every one of the major pharmaceutical companies has announced a partnership with at least one such firm. Only a few AI-discovered drugs are actually in the human-testing pipeline, however, and none has begun phase 3 human trials, the gold-standard test for experimental drugs. Saha concedes that it will be several years before he can say for sure whether the company's hit rates will go up as a result of the AI prediction rate of CYP450 inhibition. For all the hype in the industry, it is far from certain that early results will translate to more and better drugs.

SIFTING THROUGH MILLIONS OF MOLECULES

EMERGING AI PROGRAMS are not exactly a revolutionary update in the drug industry, which has for some time been building sophisticated analytical solutions that aid with drug development. The rise of powerful statistical and biophysical modeling programs well over a decade ago as part of the growth of the field of bioinformatics—the quest to use computational tools to derive biological insights from large

amounts of data—led to tools that can predict the properties of molecules. But these programs have been limited by scientists' incomplete understanding of how molecules interact: they cannot tell conventional software how to find insights in data when they do not know what elements of the data are most important and how they relate to one another. Imbued with the ability to derive their own insights into which data elements matter, newer AI programs can extract better predictions for a wider range of variables.

AI tools tackle different aspects of drug discovery in several ways. Some AI companies, for example, are focusing on the problem of designing a drug that can safely and effectively work on a known target—usually a specific, well-studied protein that is associated with a disease. The goal is typically to come up with a molecule that can chemically bind to the target protein and modify it so that it no longer contributes to the disease or its symptoms. Cyclica, a Canadian firm, puts its software to work on matching the biophysical structures and biochemical properties of millions of molecules to the structures and properties of some 150,000 proteins to uncover molecules likely to bind to target proteins, as well as those to avoid.

But molecules that are good candidates as drugs still have to jump through other hoops. Those include making it through the gut into the bloodstream without being immediately broken down by the liver or metabolic processes; working in a particular organ such as the kidney without disrupting other organs; avoiding binding to and incapacitating any of the thousands of other proteins in the human body that are important to health; and breaking down and leaving the body before drug levels become potentially dangerous. Cyclica's AI software takes all those requirements into consideration. "A molecule that can interact with a protein target can usually interact with upward of 300 proteins," Cyclica's CEO Naheed Kurji says. "If you're designing a molecule, it behooves you to consider the other 299 interactions that could have disastrous effects in humans."

There is growing recognition among biomedical researchers that complex diseases such as cancer and Alzheimer's involve hundreds of proteins, and hitting just one of them is not likely to be disruptive enough. Cyclica is attempting to find individual compounds that can interact with dozens of target proteins yet avoid interacting with hundreds of other proteins, Kurji explains. Currently under development, he adds, is the incorporation of a wealth of anonymized global genetic data about variations in proteins, so that the software can specify which patients the candidate drugs would work best on. Kurji claims that together these features will eventually be able to shave five years off the typical seven-year-long time frame for bringing a candidate drug from initial identification to human trials.

Merck and Bayer are among the big pharma companies

that have announced partnerships with Cyclica. As is the case with most AI-pharma partnerships, the companies are not releasing much insight into exactly what AI-generated drug candidates may be coming out of the collaborations. But Cyclica has shared some details of its successes in identifying a key target protein linked to already FDA-approved drugs for systemic sclerosis, an autoimmune disease of the skin and other organs, as well as one linked to the Ebola virus. Each drug is already FDA-approved for the treatment of other disorders—HIV and depression, respectively—which means they both could be quickly "repurposed" for the new applications if the research continues to pan out.

Sometimes researchers identify a target protein that might play a critical role in disease but find that—as is true of about 90 percent of the proteins in the human body—not much is known about its structure and properties. With little data to go on, most machine- and deep-learning

Resources are now piling into the field. Only a few AI-discovered drugs are actually in the human-testing pipeline, however, and none has begun phase 3 human trials, the gold-standard test for experimental drugs.

programs will not be able to figure out how to "drug" the protein—that is, come up with compounds that will bind to it and meet the other criteria for safety and efficacy. A handful of AI companies are focusing on these kinds of "small data" problems, including Exscientia, which uses its software to hunt down molecules that might work with a target protein. It can produce useful insights with as few as 10 pieces of data about a protein, says the company's CEO, Andrew Hopkins, a professor of medicinal informatics at the University of Dundee in Scotland.

Exscientia's algorithms compare the limited information available about a target protein against a database of about a billion protein interactions. This step narrows down the list of possible compounds that might work and specifies what additional data would help further refine the focus. Such data might come from looking at tissue samples to learn more about how the protein behaves in the body, for example. The resulting new data are then fed into the software, which pares the list again and suggests another round of needed data. This process is repeated until the software is

ready to generate a manageable list of compounds that are favorable drug candidates for the target.

Hopkins claims that Exscientia's process can cut the time spent in discovery from 4.5 years to as little as one year, reduces discovery costs by 80 percent and results in one-fifth the number of synthesized compounds as is normally needed to produce a single winning drug. Exscientia is partnering with biotech giant Celgene in an effort to find new potential drugs for three targets.

Meanwhile an Exscientia partnership with GlaxoSmithKline has led to what the companies say is a promising molecule targeting a novel pathway to treat chronic obstructive pulmonary disease. But as with any AI company addressing drug development, Exscientia simply has not been in the game long enough to have generated enough new candidates that could have made it through to late-stage trials—a process that typically takes five to eight years. Hopkins claims one of the candidates Exscientia has identified may reach human trials as early as this year. “At the end of the day we'll be judged on the drugs we deliver,” he says.

THE NEED FOR NEW TARGETS

FINDING A MOLECULE to hit a new target is not the only major challenge in drug discovery. There is also the need to identify targets in the first place. To spot proteins that might have roles in diseases, biopharma company Berg applies AI to sift through information derived from human tissue samples. This approach aims to solve two problems that hang over most research into drug targets, according to Berg's CEO Niven R. Narain: the efforts tend to be based on a researcher's theory or hunch, which can bias the results and overly restrict the pool of candidates, and they often turn up targets that are correlated to the disease but do not ultimately prove causative, which means drugging them will not help.

Berg's approach involves plugging in every piece of data that can be wrung out of a patient's tissue samples, organ fluids and bloodwork. These extracted data include genomics, proteomics, metabolomics, lipidomics, and more—an unusually broad range to consider in a hunt for targets. Samples are taken from people with and without a particular disease and at different stages of disease progression. Living cells from the samples are exposed in the laboratory to various compounds and conditions, such as low levels of oxygen or high levels of glucose. This method produces data on corresponding changes ranging from a cell's ability to produce energy to the rigidity of its membrane.

All the data are then run through a set of deep-learning programs that search for any differences between nondisease and disease states, with an eye to eventually focusing on proteins whose presence seem to have an impact on the disease. In some cases, those proteins become candidates as targets, at which point Berg's software can start searching for compounds to drug those targets. What is more, because the software can discern when the target seems to cause disease

in only a subset of patients, it can look for distinguishing characteristics of those patients, such as certain genes. That paves the way for a precision-medicine approach, meaning patients can be tested before they take the drug to determine whether it is likely to be effective for them.

The most exciting drug to come out of Berg's work—and perhaps the most exciting to emerge from any drug-discovery-related AI effort to date—is a cancer drug called BPM31510. It recently completed a phase 2 trial for patients with advanced pancreatic cancer, which is extremely aggressive and difficult to treat. Phase 1 trials often do not indicate much about a drug's potential except whether it is dangerously toxic at a given dose, but BPM31510's phase 1 trial against other cancers provided some verification of the ability of Berg's software to predict the roughly 20 percent of patients who were likely to respond to it, as well as those who were more likely to experience adverse reactions.

Additionally, tissue-sample analysis from the trial led Berg's software to predict, counterintuitively, that the drug would work best against more aggressive cancers because it attacks mechanisms that play a larger role in those cancers. Should the drug gain approval, Berg might do a postmarket analysis of perhaps one out of 100 patients taking it, “so that we can keep improving how it's used,” Narain says.

Berg is partnering with pharma giant AstraZeneca to seek targets for Parkinson's and other neurological diseases and with Sanofi Pasteur to pursue improved flu vaccines. It is also working with the U.S. Department of Veterans Affairs and the Cleveland Clinic on targets for prostate cancer. The software has already identified mechanisms for diagnostic tests that could differentiate prostate cancer from benignly enlarged prostates, which currently is often difficult to do without surgery.

GETTING BEYOND THE HYPE

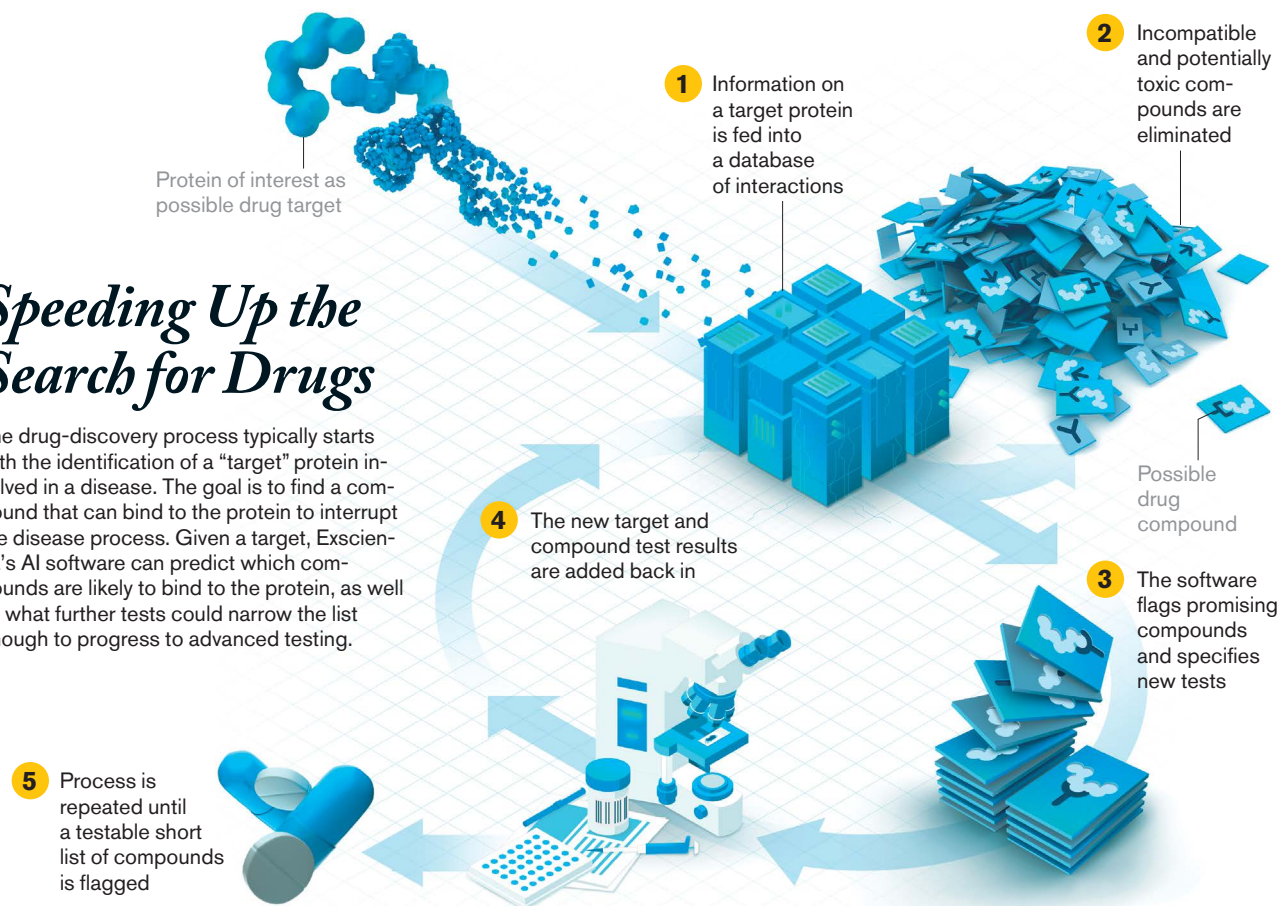
BIG PHARMA'S INTEREST in injecting these kinds of AI efforts into drug discovery can be gauged by the fact that at least 20 separate partnerships have been reported between the major companies and AI-drug-discovery tech companies. Pfizer, GlaxoSmithKline and Novartis are among the pharma companies said to have also built substantial AI expertise in-house, and it is likely that others are in the process of doing the same.

Although research executives at these companies have expressed enthusiasm for some of the early results, they are quick to admit that AI is no sure thing for the bottom line given how few new AI-aided candidates have made it to the animal-testing stage of drug development, let alone to human trials. The jury is out on whether AI will successfully make drug discovery more efficient, says Sara Kenkare-Mitra, senior vice president of development sciences at Roche subsidiary Genentech, and even if it does, “we can't yet say whether it will be an incremental improvement or an exponential leap.” If many of the drugs that result from AI efforts make it well into human testing, this question will

90
Percentage
of drug
candidates
that wash
out of devel-
opment dur-
ing human
trials, which
take place
after years
of invest-
ment have
already been
made. AI
could make
the process
more
efficient.

Speeding Up the Search for Drugs

The drug-discovery process typically starts with the identification of a “target” protein involved in a disease. The goal is to find a compound that can bind to the protein to interrupt the disease process. Given a target, Exscientia’s AI software can predict which compounds are likely to bind to the protein, as well as what further tests could narrow the list enough to progress to advanced testing.



still not be answered fully unless the drugs progress all the way through to FDA approval.

Bristol-Myers Squibb’s Saha suggests that AI-aided drugs’ rate of entry into the market is likely to remain low for some time. That rate could pick up dramatically, however, if the processes for testing and approval were streamlined to take into account the ability of machine- and deep-learning systems to more accurately predict which drugs are highly likely to be safe and effective and which patients they are best suited for. “When regulatory agencies see the same value we see in AI, the floodgates could open,” he says. “In some cases, we might be allowed to pass over animal models and go straight to human testing once we show these drugs can hit their targets with no toxicity.” But those changes are probably many years away, he admits. He adds that it is wrong to imply that AI replaces scientists and conventional research—whereas AI supports and amplifies human efforts, it still depends on humans to generate novel biological insights, set research directions and priorities, guide and validate results, and produce needed data.

The breathless hype around AI-based drug discovery might actually be damaging, Berg’s Narain says, because overpromising could lead to disappointment and backlash. “These are early days, and we need to be sober about the fact that these are tools that can help—they’re not solutions yet,” he says. Cyclica’s Kurji points the finger at AI companies that make what he says are exaggerated marketing claims, such as having reduced the many years and bil-

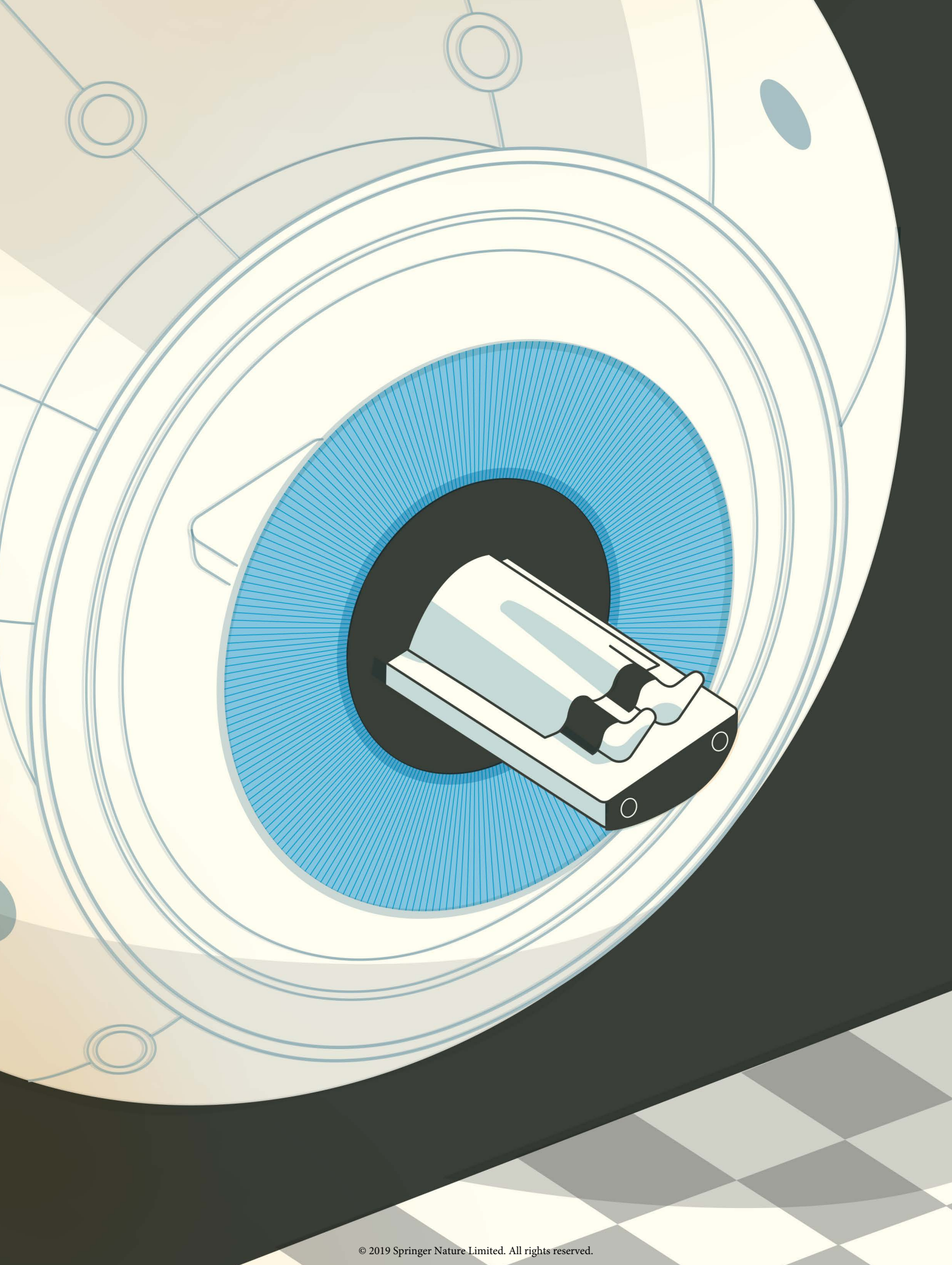
lions of dollars it takes to develop a drug to a few weeks and a few hundred thousand dollars. “It’s simply not true,” he says. “And it’s irresponsible and destructive to say so.”

But if hype hurts, Kurji insists he also knows what will give the AI-drug-discovery industry a big boost: more high-quality information to feed the various programs. “We rely on three things: data, data and more data,” he says. That sentiment is echoed by Enoch Huang, vice president of medicinal sciences at Pfizer, who says that having the right algorithm isn’t the most important factor.

The need to feed AI software with large volumes of relevant data is actually starting to change science, as researchers run more experiments specifically with the production of AI-relevant data in mind. Genentech’s Kenkare-Mitra notes that this has already happened in immunotherapy drug research. “There aren’t always enough data from the clinic to use with machine learning,” she says. “But we can [often] generate that data in vitro and feed them to the system.”

That kind of approach could lead to a virtuous cycle in drug discovery in which AI helps elucidate areas where researchers need to look for targets and drugs. Moreover, the resulting research provides larger, more relevant data sets that allow the software to point to even more fertile research avenues. “It’s not so much AI we believe in,” Kenkare-Mitra says, “as a human-AI partnership.”

David H. Freedman is a journalist who has been covering science, business and technology for more than 30 years.



Rise of Robot Radiologists

Deep-learning algorithms are peering into MRIs and x-rays with unmatched vision, but who is to blame when they make a mistake?

By Sara Reardon

WHEN REGINA BARZILAY had a routine mammogram in her early 40s, the image showed a complex array of white splotches in her breast tissue. The marks could be normal, or they could be cancerous—even the best radiologists often struggle to tell the difference. Her doctors decided the spots were not immediately worrisome. In hindsight, she says, “I already had cancer, and they didn’t see it.”

Over the next two years Barzilay underwent a second mammogram, a breast MRI and a biopsy, all of which continued to yield ambiguous or conflicting findings. Ultimately she was diagnosed with breast cancer in 2014, but the path to that diagnosis had been unbelievably frustrating. “How do you do three tests and get three different results?” she wondered.

Barzilay was treated and made a good recovery. But she remained horrified that the uncertainties of reading a mammogram could delay treatment. “I realized to what extent we are unprotected with current approaches,” she

says, so she made a career-altering decision: “I absolutely have to change it.”

A computer scientist at Massachusetts Institute of Technology, Barzilay had never studied health before. Her research used machine-learning techniques—a form of artificial intelligence—for natural-language processing. But she had been looking for a new line of research and decided to team up with radiologists to develop machine-learning algorithms that use computers’ superior visual analysis to spot subtle patterns in mammograms that the human eye might miss.

Over the next four years the team taught a computer program to analyze mammograms from about 32,000 women of different ages and races and told it which women had been diagnosed with cancer within five years of the scan. They then tested the computer’s matching abilities in 3,800 more patients. Their resulting algorithm, published last May in *Radiology*, was significantly more accurate at predicting cancer—or the absence of cancer—than practices generally used in clinics. When Barzilay’s team ran the program on her own mammograms from 2012—ones her doctor had cleared—the algorithm correctly predicted she was at a higher risk of developing breast cancer within five years than 98 percent of patients.

AI algorithms not only spot details too subtle for the human eye to see. They can also develop entirely new ways of interpreting medical images, sometimes in ways humans do not understand. The numerous researchers, start-up companies and scanner manufacturers designing AI programs hope they can improve the accuracy and timeliness of diagnoses, provide better treatment in developing countries and remote regions that lack radiologists, reveal new links between biology and disease, and even help to predict how soon a person will die.

AI applications are entering clinics at a rapid rate, and physicians have met the technology with equal parts excitement about its potential to reduce their workload and fear about losing their jobs to machines. Algorithms also raise unprecedented questions about how to regulate a machine that is constantly learning and changing and who is to blame if an algorithm gets a diagnosis wrong. Still, many physicians are excited about the promise of AI programs. “If these models can be sufficiently vetted and we can raise our level of understanding of how they work, this can help raise the level of health care for everybody,” says Matthew Lungren, a radiologist at Stanford University.

“A VERY, VERY HOT TOPIC”

THE IDEA OF using computers to read radiological scans is not new. In the 1990s radiologists started using a program called computer-assisted diagnosis (CAD) to detect breast cancer in mammograms. The technology was hailed as revolutionary, and clinics adopted it rapidly. But CAD proved to be more time-consuming and difficult to use than existing methods,

and according to some studies, clinics that used it made more errors than those that did not. The failure made many physicians dubious of computer-aided diagnostics, says Vijay Rao, a radiologist at Jefferson University in Philadelphia.

In the past decade, however, computer vision has improved by leaps and bounds—in everyday applications such as face recognition and in medicine. The advance has been largely driven by the development of deep-learning methods, in which a computer is given a set of images and then left to draw its own connections between them, ultimately developing a network of associations. In medical imaging, this might, for example, involve telling the computer which images contain cancer and setting it free to find features common to those images but absent in cancer-free images.

Development and adoption of AI technologies in radiology has spread rapidly. “Last year, at every large meeting I went to, the main theme was AI and imaging,” says Rao, past president of the Radiological Society of North America. “Clearly, this is a very, very hot topic.”

The U.S. Food and Drug Administration says that it does not keep a list of AI products that it has approved. But Eric Topol, a digital medicine researcher at the Scripps Research Institute in La Jolla, Calif., estimates that the agency is approving more than one medical imaging algorithm per month. A 2018 survey by marketing-intelligence firm Reaction Data found that 84 percent of U.S. radiology clinics had adopted or planned to adopt AI programs. The field is growing especially quickly in China, where more than 100 companies are designing AI applications for health care.

“It’s a fascinating time to be in this market,” says Elad Walach, CEO of the Tel Aviv-based start-up Aidoc. The company develops algorithms to analyze CT scans for abnormalities and move those patients to the top of a doctor’s priority list. Aidoc also tracks how often doctors use the program and how long they spend second-guessing its conclusions. “Initially they’re skeptical, but after two months they get used to it and are very trusting,” Walach says.

Saving time can be crucial to saving a patient. One recent study of chest x-rays for collapsed lungs found that radiologists flag more than 60 percent of the scans they order as high priority, which suggests that they might spend hours wading through nonserious cases before getting to those that are actually urgent. “Every doctor I talk to has a story where they lost a patient because of a collapsed lung,” says Karley Yoder, vice president and general manager of AI at Boston-based GE Healthcare, one of the leading manufacturers of medical imaging equipment. Last September the FDA approved a set of AI tools that will now come embedded in GE scanners, automatically flagging the most urgent cases.

Because they can process massive amounts of data, computers can perform analytical tasks that are beyond human capability. Google, for instance, is using its computing power to develop AI algorithms that construct two-dimen-

sional CT images of lungs into a three-dimensional lung and look at the entire structure to determine whether cancer is present. Radiologists, in contrast, have to look at these images individually and attempt to reconstruct them in their heads. Another Google algorithm can do something radiologists cannot do at all: determine patients’ risk of cardiovascular disease by looking at a scan of their retinas, picking up on subtle changes related to blood pressure, cholesterol, smoking history and aging. “There’s potential signal there beyond what was known before,” says Google product manager Daniel Tse.

THE BLACK BOX PROBLEM

AI PROGRAMS COULD END UP revealing entirely new links between biological features and patient outcomes. A 2019 paper in *JAMA Network Open* described a deep-learning algorithm trained on more than 85,000 chest x-rays from people enrolled in two large clinical trials that had tracked them for more than 12 years. The algorithm scored each patient’s risk of dying during this period. The researchers found that 53 percent of the people the AI put into a high-risk category died within 12 years, as opposed to 4 percent in the low-risk category. The algorithm did not have information on who died or on the cause of death. The lead investigator, radiologist Michael Lu of Massachusetts General Hospital, says that the algorithm could be a helpful tool for assessing patient health if combined with a physician’s assessment and other data such as genetics.

To understand how the algorithm worked, the researchers identified the parts of images that it used to make its calculations. Some, such as waist circumference and the structure of women’s breasts, made sense because these areas can hint at known risk factors for certain diseases. But the algorithm also looked at the region under patients’ shoulder blades, which has no known medical significance. Lu guesses that flexibility might be one predictor of a shorter life span. Taking a chest x-ray often requires patients to hug the machine, and less healthy people who cannot put their arms all the way around it might position their shoulders in a different way. “They’re not things I would have thought of *de novo* and might not understand,” Lu says.

The disconnect between the way computers and humans think is known as the black box problem: the idea that a computer brain operates in an obscured space that is inaccessible to humans. Experts differ on whether this presents a problem in medical imaging. On the one hand, if an algorithm consistently improves doctors’ performance and patients’ health, doctors do not need to know how it works. After all, researchers still do not fully understand the mechanisms of many drugs such as lithium, which has been used to treat depression since the 1950s. “Maybe we shouldn’t be so fixated, because the way humans work in medicine is about as black box as you can get,” Topol says. “Do we hold machines to a higher standard?”

84
Percentage
of U.S.
radiology
clinics that
have
adopted
or plan to
adopt AI
programs,
according
to a 2018
survey.

“AI won’t replace radiologists, but radiologists who use AI will replace radiologists who don’t.”
—Curtis Langlotz, Stanford University

Still, there is no denying that the black box presents ample opportunity for human-AI misunderstanding. For instance, researchers at the Icahn School of Medicine at Mount Sinai were deeply puzzled by a discrepancy in the performance of a deep-learning algorithm they had developed to identify pneumonia in lung x-rays. It performed with greater than 90 percent accuracy on x-rays produced at Mount Sinai but was far less accurate with scans from other institutions. They eventually figured out that instead of just analyzing the images, the algorithm was also factoring in the odds of a positive finding based on how common pneumonia was at each institution—not something they expected or wanted the program to do.

Confounding factors like these worry Samuel Finlayson, who studies biomedical applications of machine learning at Harvard Medical School. He notes that data sets on which AI is trained can be biased in ways that developers fail to consider. An image taken in an emergency room or one taken in the middle of the night may be more likely to show a sick person than one taken during a routine examination, for instance. An algorithm could also learn to look at scars or medical device implants that indicate a previous health problem and decide that people without these marks did not have the condition. Even the way that institutions label their images can confuse an AI algorithm and prevent the model from functioning well in another institution with a different labeling system. “If you naively train [an algorithm] at a hospital from one location, one time, and one population group, you’re unaware of all the thousands of little factors that models are taking into account. If any of those change, you can be in for a world of hurt,” Finlayson warns.

The solution, Finlayson says, is to train an algorithm with data from many locations and in diverse patient populations, then test it prospectively—without any modifications—in a new patient population. But very few algorithms have been tested this way. According to Topol’s recent *Nature Medicine* review, among dozens of studies claiming an AI performs better than radiologists, only a handful were tested in populations that were different from the population where they were developed. “Algorithms are very, very delicate,” says Cynthia Rudin, a computer scientist at Duke University. “If you try to use one outside the training set [of images], it doesn’t always work.”

As researchers become aware of this problem, more prospective studies in novel settings could be on the horizon. Barzilay’s team recently finished testing its mammogram AI on 10,000 scans from the Karolinska Institute in Sweden and found that it performed just as well there as it did in Massachusetts. The group is now working with hospitals in Taiwan and Detroit to test it in more diverse patient groups. The team found that current standards for assessing breast cancer risk are much less accurate in African-American women, Barzilay says, because those standards were devel-

oped mostly using scans from white women: “I think we really are in a position to revamp this sad state of affairs.”

LEGAL TERRA INCOGNITA

EVEN IF THE AI’S conclusions are medically relevant, the black box still presents a number of concerns from a legal perspective. If an AI gets a diagnosis wrong, it can be hard to determine whether the doctor or the program is at fault. “Lots of bad things happen in health care, and you don’t necessarily know why the bad things happened,” says Nicholas Price, a health law expert at the University of Michigan. If an AI system leads a physician to make an incorrect diagnosis, the physician may not be able to explain why and the company’s data on the test’s methodology are likely to be a closely guarded trade secret.

Medical AI systems are still too new to have been challenged in medical malpractice lawsuits, so it is unclear how courts will determine responsibility and what kind of transparency should be required.

The tendency to build black box algorithms frustrates Rudin. The problem comes from the fact that most medical algorithms are built by adapting deep-learning tools developed for other types of image analysis. “There’s no reason you can’t build a robot that can explain itself,” she insists. But it is exponentially harder to build a transparent algorithm from scratch than to repurpose an existing black box algorithm to look at medical data. That is why Rudin suspects most researchers let an algorithm run and then try to understand later how it came to its conclusion.

Rudin is developing transparent AI algorithms that analyze mammograms for suspected tumors and constantly inform researchers what they are doing. But her research has been stymied by the lack of available images on which to train the algorithm. The images that are publicly available tend to be poorly labeled or taken with old machines that are no longer in use, Rudin says, and without enormous, diverse data sets, algorithms tend to pick up confounding factors.

Black boxes, along with an AI algorithm’s ability to learn from experience, also present challenges to regulators. Unlike a drug, which will always work in the same way, machine-learning algorithms change and improve over time as they gain access to more patient data. Because the algorithm draws meaning from so many kinds of input, seemingly innocuous changes such as a new IT system at a hospital could suddenly ruin the AI program. “Machines can get sick just like humans get sick, and they can be in-

“You can’t trust an algorithm when someone’s life is on the line.” —*Eric Topol, Scripps Research*

fectured with malware,” Topol says. “You can’t trust an algorithm when you have someone’s life on the line.”

Last April the FDA proposed a set of guidelines to manage algorithms that evolve over time. Among them is an expectation that producers keep an eye on how their algorithms are changing to ensure they continue to work as designed and asking them to notify the agency if they see unexpected changes that might prompt reevaluation. The agency is also developing best manufacturing practices and may require companies to spell out their expectations for how algorithms might change and a protocol for how to manage those changes. “We need to understand that not one size fits all,” says Bakul Patel, director of digital health at the FDA.

WILL MACHINES REPLACE DOCTORS?

THE LIMITATIONS OF AI should reassure radiologists who worry about machines taking their jobs. In 2012 technology venture capitalist and Sun Microsystems co-founder Vinod Khosla horrified a medical audience by predicting that algorithms would replace 80 percent of doctors, and more recently he claimed that radiologists still practicing in 10 years will be “killing patients.” Such remarks caused panic and backlash in the radiology field, Rao says. “I think the hype is creating a lot of expectations.”

But that concern has also had real impacts. In 2015 only 86 percent of radiology resident positions in the U.S. were filled, compared with 94 percent the previous year, although those numbers have improved over the past several years. And according to a 2018 survey of 322 Canadian medical students, 68 percent believed AI would reduce the demand for radiologists.

Still, most experts and AI manufacturers doubt AI will be replacing doctors any time soon. “AI solutions are becoming very good at doing one thing very well,” Walach says. But because human biology is complex, he says, “you typically have to have humans who do more than one thing really well.” In other words, even if an algorithm is better at diagnosing a particular problem, combining it with a physician’s experience and knowledge of the patient’s individual story will lead to a better outcome.

An AI that can do a single task well could free radiologists from drudgework, allowing them more time to interact with patients. “They could come out of the basement, which is where they live in the dark,” Topol says. “What we need in medicine is more interhuman contact and bonding.”

Still, Rao and others believe that the tools and training that radiologists receive, including their day-to-day work, will change drastically over the coming years as a result of artificial-intelligence algorithms. “AI won’t replace radiol-

ogists, but radiologists who use AI will replace radiologists who don’t,” says Curtis Langlotz, a radiologist at Stanford.

There are some exceptions, however. In 2018 the FDA approved the first algorithm that can make a medical decision without the need for a physician to look at the image. The program, developed by IDx Technology in Coralville, Iowa, looks at retinal images to detect diabetic retinopathy and is 87 percent accurate, according to the company’s data. IDx chief executive officer Michael Abramoff says that because no doctor is involved, the company has assumed legal liability for any medical errors.

In the short term, AI algorithms are more likely to assist doctors than replace them. For instance, physicians working in developing countries might not have access to the same kinds of scanners as a major medical institution in the U.S. or Europe or trained radiologists who can interpret scans. As medicine becomes more specialized and dependent on image analysis, the gap between the standard of care provided in wealthier and poorer areas is growing, Lungren says. Running an algorithm can be a cheap way to close that gap and may even be done on a mobile phone.

Lungren’s group is developing a tool that allows doctors to take cell-phone pictures of an x-ray film—not the digital scans that are standard in wealthy nations—and run an algorithm on the photographs that detects problems such as tuberculosis. “It’s not replacing anybody,” he says—many developing countries have no radiologists in the first place. “We’re augmenting nonradiologists to bring expertise to their fingertips.”

Another short-term goal of AI could be to examine medical records to determine whether a patient needs a scan in the first place, Rao says. Many medical economists believe that imaging is overused—more than 80 million CT scans are performed every year in the U.S. alone. Although this abundance of data is helpful to researchers using it to train algorithms, scans are extraordinarily costly and can expose patients to unnecessary amounts of radiation. Similarly, Langlotz adds that algorithms could one day analyze images while a patient is still in the scanner and predict the final outcome, thus reducing the amount of time and radiation exposure required to get a good image.

Ultimately, Barzilay says, AI will be most useful when it serves as a sharp-eyed partner in tackling problems that doctors cannot detect and solve alone. “If there were a convenient and describable pattern,” she notes, “humans would already be able to do it.” She knows firsthand that, too often, this is not the case.

Sara Reardon is a freelance journalist based in Bozeman, Mont. She is a former staff reporter at *Nature*, *New Scientist* and *Science* and has a master’s degree in molecular biology.



Can AI Fix Medical Records?

Digitized patient charts were supposed to revolutionize medical practice. Artificial intelligence could help unlock their potential

By Cassandra Willyard

A YOUNG MAN, let's call him Roger, arrives at the emergency department complaining of belly pain and nausea. A physical exam reveals that the pain is focused in the lower right portion of his abdomen. The doctor worries that it could be appendicitis. But by the time the imaging results come back, Roger is feeling better, and the scan shows that his appendix appears normal. The doctor turns to the computer to prescribe two medications, one for nausea and Tylenol for pain, before discharging him.

This is one of the fictitious scenarios presented to 55 physicians around the country as part of a study to look at the usability of electronic health records (EHRs). To prescribe medications, a doctor has to locate them in the EHR system. At one hospital a simple search for Tylenol brings up a list of more than 80 options. Roger is a 26-year-old man, but the list includes Tylenol for children and infants, as well as Tylenol for menstrual cramps. The doctor tries to winnow the list by typing the desired dose—500 milligrams—into the search window, but now she gets zero hits. So she returns to the main list and finally selects the 68th option—Tylenol Extra Strength (500 mg), the most commonly prescribed dose of Tylenol. What should have been a simple task has taken precious minutes and far more brainpower than it deserved. This is just one example of the countless agonizing frustrations that physicians deal with every day when they use EHRs.

These EHRs—digital versions of the paper charts in which doctors used to record patients' visits, laboratory results and other important medical information—were supposed to transform the practice of medicine. The Health Information Technology for Economic and Clinical Health (HITECH) Act, passed in 2009, has provided \$36 billion in financial incentives to drive hospitals and clinics to transition from paper charts to EHRs. Then president Barack Obama said the shift would “cut waste, eliminate red tape and reduce the need to repeat expensive medical tests.” He added that it would “save lives by reducing the deadly but preventable medical errors that pervade our health care system.”

When HITECH was adopted, 48 percent of physicians used EHRs. By 2017 that number had climbed to 85 percent, but the transformative power of EHRs has yet to be realized. Physicians complain about clunky interfaces and time-consuming data entry. Polls suggest that they spend more time interacting with a patient's file than with the actual patient. As a result, burnout is on the rise. Even Obama observed that the rollout did not go as planned. “It's proven to be harder than we expected,” he told Vox in 2017.

Yet EHRs do have the potential to deliver insights and efficiencies, according to physicians and data scientists. Artificial intelligence in the form of machine learning—which allows computers to identify patterns in data and draw conclusions on their own—might be able to help overcome the obstacles encountered with EHRs and unlock their potential for making predictions and improving patient care.

DIGITAL DEBACLE

IN 2016 the American Medical Association teamed up with MedStar Health, a health care organization that operates 10 hospitals in the Baltimore-Washington area, to examine the usability of two of the largest EHR systems, developed by Cerner, based in North Kansas City, Mo., and Epic, based in Verona, Wis., respectively. Together these two companies account for 54 percent of the acute care hospital market. The

team recruited emergency physicians at four hospitals and gave them fictitious patient data and six scenarios, including the one about Roger, who presented with what seemed like appendicitis. These scenarios asked the physicians to perform common duties such as prescribing medications and ordering tests. The researchers assessed how long it took the physicians to complete each task, how many clicks were required and how accurately they performed.

What they found was disheartening. The time and the number of clicks required varied widely from site to site and even between sites using the same system. And some tasks, such as tapering the dose of a steroid, proved exceptionally tricky across the board. Physicians had to manually calculate the taper doses, which took anywhere from two to three minutes and required 20 to 42 clicks. These design flaws were not benign. The physicians often made dosage mistakes. At one site the error rate reached 50 percent. “We've seen patients being harmed and even patients dying because of errors or issues that arise from usability of the system,” says Raj Ratwani, director of MedStar Health's National Center for Human Factors in Healthcare.

But clunky interfaces are just part of the problem with EHRs. Another stumbling block is that information still does not flow easily between providers. The system lacks “the ability to seamlessly and automatically deliver data when and where it is needed under a trusted network without political, technical, or financial blocking,” according to a 2018 report from the National Academy of Medicine. If a patient changes doctors, visits urgent care or moves across the country, her records might or might not follow. “Connected care is the goal; disconnected care is the reality,” the authors wrote.

In March 2018 the Harris Poll conducted an online survey on behalf of Stanford Medicine that examined physicians' attitudes about EHRs. The results were sobering. Doctors reported spending, on average, about half an hour on each patient. More than 60 percent of that time was spent interacting with the patient's EHR. Half of office-based primary care physicians think using an EHR actually diminishes their clinical effectiveness. Isaac Kohane, a computer scientist and chair of the department of biomedical informatics at Harvard Medical School, puts it bluntly: “Medical records suck.”

Yet despite the considerable drawbacks of existing EHR systems, most physicians agree that electronic records are a vast improvement over paper charts. Getting patients' data digitized means that they are now accessible for analysis using the power of AI. “There's huge potential to use artificial intelligence and machine learning to develop predictive models and better understand health outcomes,” Ratwani says. “I think that's absolutely the future.”

It is already happening to some extent. In 2015 Epic began offering its clients machine-learning models. To develop these models, computer scientists start with algorithms and train them using real-world examples with known outcomes. For example, if the goal is to predict which

“Health data is like crude oil. It is useless unless it is refined.” —Leo Anthony Celi, M.I.T. Laboratory for Computational Physiology

patients are at greatest risk of developing the life-threatening blood condition known as sepsis, which is caused by infection, the algorithm might incorporate data routinely collected in the intensive care unit, such as blood pressure, pulse and temperature. The better the data, the better the model will perform.

Epic now has a library of models that its customers can purchase. “We have over 300 organizations either running or implementing models from the library today,” says Seth Hain, director of analytics and machine learning at Epic. The company’s sepsis-prediction model, which scans patients’ information every 15 minutes and monitors more than 80 variables, is one of its most popular. The North Oaks Health System in Hammond, La., implemented the model in 2017. If a patient’s score reaches a certain threshold, the physicians receive a warning, which signals them to monitor the patient more closely and provide antibiotics if needed. Since the health system implemented the model, mortality caused by sepsis has fallen by 18 percent.

But building and implementing these kinds of models is trickier than it might first appear. Most rely solely on an EHR’s structured data—data that are collected and formatted in the same way. Those data might consist of a blood-pressure reading, lab results, a diagnosis or a drug allergy. But EHRs include a wide variety of unstructured data, too, such as a clinician’s notes about a visit, e-mails and x-ray images. “There is information there, but it’s really hard for a computer to extract it,” says Finale Doshi-Velez, a computer scientist at Harvard University. Ignoring this free text means losing valuable information, such as whether the patient has improved. “There isn’t really a code for doing better,” she says. Moreover, Ratwani points out that because of poor usability, data often end up in the wrong spot. For example, a strawberry allergy might end up documented in the clinical notes rather than being listed in the allergies box. In such cases, a model that looks for allergies only in the allergy section of the EHR “is built off of inaccurate data,” he adds. “That is probably one of the biggest challenges we’re facing right now.”

Leo Anthony Celi, an intensive care specialist and clinical research director at the Massachusetts Institute of Technology’s Laboratory for Computational Physiology, agrees. Most of the data found in EHRs are not ready to be fed into an algorithm. A massive amount of curation has to occur first. For example, say you want to design an algorithm to help patients in the intensive care unit avoid low blood glucose, a common problem. That sounds straightforward, Celi says. But it turns out that blood sugar is measured in different ways, with blood drawn from either a finger prick or a vein. Insulin is administered in different ways, too. When Celi and his colleagues examined all the data on insulin and blood sugar from patients at one hospital, “there were literally thousands of different ways they were entered in the EHR.” These data have to be manually sorted and clustered by type before one can even design an algorithm. “Health data is like crude oil,” Celi says. “It is useless unless it is refined.”

AN INTELLIGENT FIX

THE CURRENT PITFALLS of EHRs hamper efforts to use artificial intelligence to glean important insights, but AI might itself provide a possible solution. One of the main drawbacks of the existing EHR

systems, doctors say, is the time it takes to document a visit—everything from the patient’s complaint to the physician’s analysis and recommendation. Many physicians believe that much of the therapeutic value of a doctor visit is in the interactions, Kohane says. But EHRs have “literally taken the doctor from facing the patient to facing the computer.” Doctors have to type up their narrative of the visit, but they also enter much of the same information when they order lab tests, prescribe medications and enter billing codes, says Paul Brient, chief product officer at athenahealth, another EHR vendor. This kind of duplicate work contributes to physician frustration and burnout.

As a stopgap measure, some hospitals now have scribes sit in on appointments to document the visit while the physician interacts with the patient. But several companies are working on digital scribes, machine-learning algorithms that can take a conversation between a doctor and a patient, parse the text and use it to fill in the relevant information in the patient’s EHR.

Indeed, some such systems are already available. In 2017 Saykara, a Seattle-based start-up, launched a virtual assistant named Kara. The iOS app uses machine learning, voice recognition and language processing to capture conversations between patients and physicians and turn them into notes, diagnoses and orders in the EHR. Previous versions of the app required prompts from the physician—much like Apple’s Siri—but the current version can be put in “ambient mode,” in which it simply listens to the entire conversation and then selects the relevant information. EHRs turned physicians into data-entry clerks, Kohane says. But apps like Kara could serve as intelligent, knowledgeable co-workers. And Saykara is just one of a host of start-ups developing such tools. Athenahealth’s latest mobile app allows physicians to dictate their documentation. The app then translates that text into the appropriate billing and diagnostic codes. But “it’s not perfect by any stretch of the imagination,” Brient says. The physician still has to check for errors. The app does reduce the workload, however. The systems that Robert Wachter, chair of the department of medicine at the University of California, San Francisco, has seen are “probably not quite ready for prime time,” he says, but they should be in a couple of years.

Artificial intelligence might also help clinicians make better, more sophisticated decisions. “We think of the decision support in a computer system as an alert,” says Jacob Reider, a physician and CEO at Alliance for Better Health, a New York-based health care system that works to improve the health of communities. That alert might be a box that pops up to warn of a drug allergy. But a more sophisticated system might list the likelihood of a side effect with drug option A versus drug option B and provide a cost comparison. From a technological standpoint, developing such a feature is “no different from Amazon putting an advertisement or making you aware of a purchasing opportunity,” he says.

Wachter sees at least one encouraging sign that progress is coming. In the past few years the behemoths of the tech world—Google, Amazon, Microsoft—have developed a strong interest in health care. Google, for example, partnered with researchers from U.C.S.F., Stanford University and the University of Chicago to develop models aimed at predicting events relevant to hospitalized patients, such as mortality and unexpected readmission.

To deal with the messy data problem, the researchers first translated data from two EHR systems into a standardized format called Fast Healthcare Interoperability Resources, or FHIR (pronounced “fire”). Then, rather than hand-selecting a set of variables such as blood pressure and heart rate, they had the model read patients’ entire charts as they unfolded over time up until the point of hospitalization. The data unspooled into a total of 46,864,534,945 data points, including clinical notes. “What’s interesting about that approach is every single prediction uses the exact same data to make the prediction,” says Alvin Rajkomar, a physician and AI researcher at Google who led the effort. That element both simplifies data entry and enhances performance.

But the involvement of massive corporations also raises serious privacy concerns. In mid-November 2019 the *Wall Street Journal* reported that Google, through a partnership with Ascension, the country’s second-largest health care system, had gained access to the records of tens of millions of people without their knowledge or consent. The company planned to use the data to develop machine-learning tools to make it easier for doctors to access patient data.

This type of data sharing is not unprecedented or illegal. Tariq Shaukat, Google Cloud’s president of industry products and solutions, wrote that the data “cannot be used for any other purpose than for providing these services we’re offering under the agreement, and patient data cannot and will not be combined with any Google consumer data.” But those assurances did not stop the Department of Health and Human Services from opening an inquiry to determine whether Google/Ascension complied with Health Insurance Portability and Accountability Act regulations. As of press time, the inquiry was ongoing.

But privacy concerns should not halt the quest for better, smarter, more responsive electronic health records, according to Reider. There are ways to develop these systems that maintain privacy and security, he says.

Ultimately real transformation of medical practice may require an entirely new kind of EHR, one that is not simply a digital file folder. All the major EHRs are built on top of database-type architecture that is 20 to 30 years old, Reider observes. “It’s rows and columns of information.” He likens these systems to the software used to record inventory at a brick-and-mortar bookstore: “It would know which books it bought, and it would know which books it sold.” Now envision how Amazon uses algorithms to predict what a customer might buy tomorrow and to anticipate demand. “They’ve engineered their systems so that they can learn in this way, and then they can autonomously take action,” Reider says. Health care needs the same kind of transformative leap.

.....
Cassandra Willyard is a science writer based in Madison, Wis.

Wiring Minds

Successfully applying AI to biomedicine requires innovators trained in contrasting cultures

By Amit Kaushal and Russ B. Altman

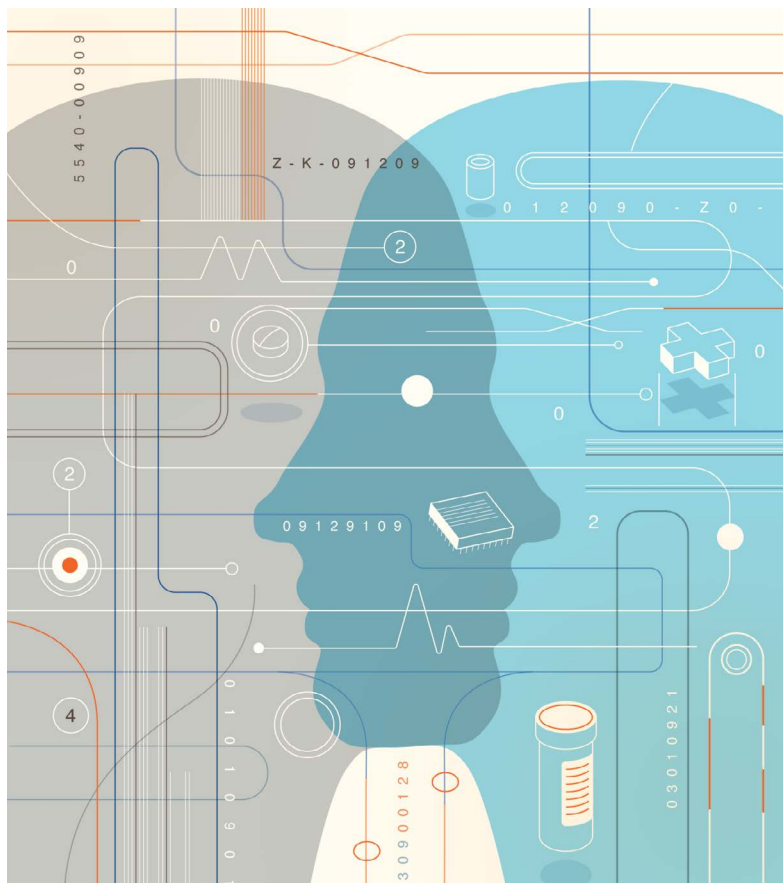
From the popular press to the largest health care conferences, promises of artificial intelligence revolutionizing biomedicine are ubiquitous. It often seems as if we are on the cusp of AI systems that can remotely identify a person about to get sick, make a diagnosis (no doctor needed!), select a custom AI-designed pharmaceutical and deliver it to the patient just in time—in an AI-powered self-driving car, of course.

If indeed this is the future, we are far from reaching it. To be sure, the pace of change has been rapid. Deep learning—the fast-growing subfield of AI that enables machines to diagnose pneumonia from chest x-rays or predict health deterioration from medical records—was unfamiliar even to most computer scientists a decade ago. And we do not know what evolutionary or revolutionary advances will drive AI in the coming decades. What we do know is that the success of biomedical AI depends not just on developing the technology but also on developing the people behind it.

Translating algorithmic advances to biomedical breakthroughs requires critically considering both realms of knowledge and endeavor on many levels. What, for example, are the true capabilities of a new technology, and what is simply hype? What problems in biomedicine are most likely to benefit from emerging computational capabilities? And how do we go from an interesting biomedical application of a new technology to the implementation of systems that actually improve human health? These challenging, multifaceted questions will need to be answered by interdisciplinary teams. The teams will require experts in AI, experts in biology and medicine, and, most important, leaders who can motivate and guide individuals with such diverse talents.

Unlike some domains in which AI has been applied, in biomedicine the consequences of failure are weighty. For a social media company, an AI model that is ineffective at increasing ad clicks can be detected and rolled back the same day. When it comes to medicine, however, human lives are at stake. Inadequately informed uses of AI can lead to obvious harm, such as inaccurate diagnostic or therapeutic recommendations, but also to more insidious failures, such as an algorithm that gives racially biased recommendations because it was trained with subtly biased data. Given the complexities of biomedicine and the inscrutable nature of many AI algorithms, it might be years before such a flaw is uncovered. Group leaders—whether in academia, pharmaceutical laboratories or start-ups—must not only understand the technical and scientific issues but also anticipate and articulate the potential risks, benefits and implications of the projects they undertake.

We need men and women who can build AI systems in med-



intuitions are still forming? The difference would be like that between learning a second language as an adult and growing up in a bilingual household: fluency is second nature for early starters.

In 2001 we launched an engineering major at Stanford University to enable undergraduates to learn computer science and statistics in the context of biology and medicine. The program creates graduates with a bachelor of science degree who have already wrestled intensively with the challenges of applying computational tools to hard problems in biomedicine. Our students take biology with premedical students and computer science with classmates who will work in Silicon Valley, and each completes a two- or three-quarter-long research project during his or her time at Stanford. They acquire knowledge with breadth across the biomedical and technical fields and depth in a narrower application area. At least one course on the societal and ethical implications of technology is also required.

icine that improve care. It is relatively easy to generate excitement by solving the technical aspects of a problem, but making those advances useful often involves wrestling with the complex interplay of regulatory, economic and workflow issues in health care systems. Successful leaders benefit from deep knowledge and intuition in both the AI and the biomedical domains. But we face a critical shortage of such versatile individuals.

Tackling this gap is crucial to ensuring the long-term success of biomedical AI. A primary challenge is the length of study required in these disciplines, but a greater one is training students in two realms that could hardly be more different in their approaches to problem-solving. Computer science involves the quantitative rigor of mathematics, statistics and engineering, whereas biology is underpinned by the haphazard products of evolution. Properties of living things are, literally and figuratively, organic. We seek students with the intellectual flexibility and passion to undergo lengthy training in both these contrasting cultures. Are we asking for the impossible?

These individuals do exist, and their numbers are growing. The first approach to their training is to identify individuals who already have a deep background in either biomedical or computational science and then help them become skilled in the other area. Graduate programs (M.S., Ph.D. and M.D./Ph.D.) in biomedical informatics have filled this role since the early 1980s. These programs attract diverse students and have grown to include disciplines that go by various names: computational biology, bioinformatics, clinical informatics, biomedical data science, and so on. All are concerned with different applications of computer science to biomedicine.

But what about training students at the intersection of these disciplines even earlier in their careers—while their intellectual

After almost two decades of training biomedical-computation undergraduates, we can say that the model works. Many of our graduates have gone on to careers in academia, clinical medicine, start-up companies (both in and outside of the biology field), large companies, law firms, venture capital, and elsewhere. And the major has consistently drawn a 50–50 balance of men and women—true for only a minority of quantitatively intensive engineering majors.

For most, the major has shaped their professional identity: they are not “AI people doing bio” or “Bio people doing AI.” Instead both of these intellectual traditions reside comfortably within their minds, each informing their understanding of the other. Whereas it is impossible to learn the entirety of biomedicine and computer science in just four years (or even in 40), these people move freely between the cultures of biology and computer science and have already learned to apply deep technical skills to the hardest societal challenges in biology and human health.

In addition to graduate programs, the development of a robust set of undergraduate programs at the interface of biomedicine and computation could give students who are in a formative period of their education the ability to move fluidly between these very different disciplines. Such programs would accelerate the emergence of the workforce required for appropriate use of AI to advance biology and health care.

Amit Kaushal is a clinical assistant professor of medicine and an adjunct professor of bioengineering at Stanford University.

Russ B. Altman is a professor of bioengineering, genetics, medicine and biomedical data science at Stanford University.



CHRISTIAN VINCES/GETTY

Universities in Peru's capital, Lima, are collaborating with teams studying malaria in the Amazon.

PERU HAS PROMISE

To compete globally, protect its natural resources and move its economy forward, Peru will need to take advantage of scientific opportunities in plain sight.

By Aleszu Bajak

In 2017, Gabriel Carrasco-Escobar started mapping mosquito breeding sites in the Peruvian Amazon. The doctoral student at the University of California, San Diego, was hoping to collect data to be used to fight malaria, a disease that is endemic in his home country of Peru.

Carrasco-Escobar wanted to help public-health authorities in the northeastern region of Loreto, to be more proactive about combating a mosquito-borne illness that

infected more than 64,000 Peruvians in 2014. By mapping mosquito habitats, he could help the authorities to predict where malaria might spread in the future, he says, “instead of being reactive and trying to control a malaria outbreak after it’s already been reported, like firefighters putting out a blaze”.

If Carrasco-Escobar and his team could get a view of the forest from above, they could identify bodies of water that might harbour mosquito larvae, he reasoned, and then use

that information to divert mosquito-control teams to those areas before an epidemic hit.

He knew satellite imagery wouldn’t be enough because the rainforest is predisposed to cloud cover. So he sought help from researchers at Cayetano Heredia University in Lima, enlisted collaborators across the United States and gained funding from the World Health Organization. With that support, he fitted out inexpensive drones with the US\$15,000 cameras he needed.

He and his team then set out into the Amazon to collect data. But soon, like countless explorers who had trudged into the jungle before him, Carrasco-Escobar learnt that the elements always win in the end. “Eventually the Amazon beats your technology. We lost a drone to the jungle. It felt like we had lost a member of our field team,” he says, with the kind of nervous laugh that recognizes how much money was lost. “It crashed into a tree and we couldn’t find it.”

These challenges — doing remote fieldwork, guiding a constellation of collaborators, tracking down international funding, confronting a major logistical hurdle that

slows the science down – are probably familiar to many scientists, especially those working in Peru. And for most Peruvian researchers, the challenges are symptomatic of systemic issues in which scientific bureaucracies, a lack of government support and an outdated education system seem to have combined to make science difficult to undertake.

That's lamentable, researchers say, because of how much the country has to offer science – Peru supports research into tropical glaciers in the high Andes, ecosystems fed by the Humboldt Current in the Pacific Ocean, childhood development and infectious diseases and theoretical physics, among myriad other fields.

To ensure they have a role in their country's development, Peruvian researchers will have to be creative about taking advantage of the scientific opportunities Peru has to offer.

"There's enormous potential in Peruvian scientists," says Gisella Orjeda, the former president of CONCYTEC, Peru's National Council of Science, Technology and Technological Innovation – if only they can work out how to harness the opportunities in plain sight.

Global collaboration

In 2018, Peru ranked seventh in Latin America in terms of the number of papers it published, and it is heavily reliant on international collaboration, as Carrasco-Escobar's malaria study exemplifies. Roughly two-thirds of scientific studies published from Peru since 2003 listed foreign collaborators, according to data from the research-outputs database Dimensions.ai (Dimensions is part of Digital Science, a firm operated by the Holtzbrinck Publishing Group, which has a share in *Nature's* publisher, Springer Nature). That's in line with the calculations of Félix de Moya Anegón, a bibliometrics researcher at the University of

Granada in Spain and founder of the Scimago Lab, which studies publishing output.

"Only 20% of Peru's international scientific collaborations are led by Peruvians," says de Moya Anegón, who analysed how many Peruvian scientists are corresponding authors on studies that took place in Peru. "This creates a dependency on outside science and a low amount of leadership from Peruvian science."

But Alberto Gago, a high-energy particle physicist originally from Peru who collaborates with teams at Fermilab, in Batavia, Illinois, the Massachusetts Institute of Technology, in Cambridge and CERN, Europe's particle-physics laboratory near Geneva, Switzerland, says pulling in international partners is essential to building a career as a Peruvian

"Fundamentally, we are an exporter of scientists to the world."

scientist. "You have limitations if you're alone," he says. "You're limited in the number of things you can study."

He credits his graduate work at Fermilab and Brazil's University of São Paulo for affording him the right mix of theory and experimentation to keep him plugged into world-class particle physics from South America. "If I didn't have that profile, Peru wouldn't be able to have a seat at the table."

Gago also praises his institution, the Pontifical Catholic University of Peru (PCUP) in Lima, for facilitating those collaborations by financially supporting his research group rather than letting it run exclusively on grant funding – an arrangement that Gago says is rare in Peru. The PCUP also built him a distributed computer system that mines idle time on up to

600 computers across campus to run his analysis. Without that high-throughput computing system, which his university dubbed LEGION, many of his projects wouldn't have been possible, he admits.

"If it wasn't for [the PCUP], we wouldn't be collaborating with CERN or with Fermilab," says Gago, who estimates he's received \$70,000 from the university for research assistants since 2002. "The strong investment has been from the university as well as international funds. But this shouldn't just come from your institution. It should come from the state."

That's a common refrain among scientists in Peru. There is little investment from the government into science – the country invests only 0.12% of its gross domestic product into science and technology, compared with 0.36% in Chile, 1.27% in Brazil and 2.8% in the United States (see 'By the numbers'). This lack of support for learning and undertaking science in the country has led to an exodus of researchers.

"Latin America has great researchers but, fundamentally, we are an exporter of scientists to the world," says Orjeda. "I can't explain how it's possible that Peru has always abandoned science."

Education reform

In 2012, when Orjeda first took the helm of CONCYTEC, her budget was only \$5 million. Divided over the whole country, that's almost nothing, she says. "There was no budget, no vision, no tools to close the gaps of science in Peru." Over her five-year tenure, the agency's funding grew almost eight-fold and was used to fund local research, create scholarships for Peruvians to do graduate work at the world's top universities and establish a repatriation programme to entice them to return.

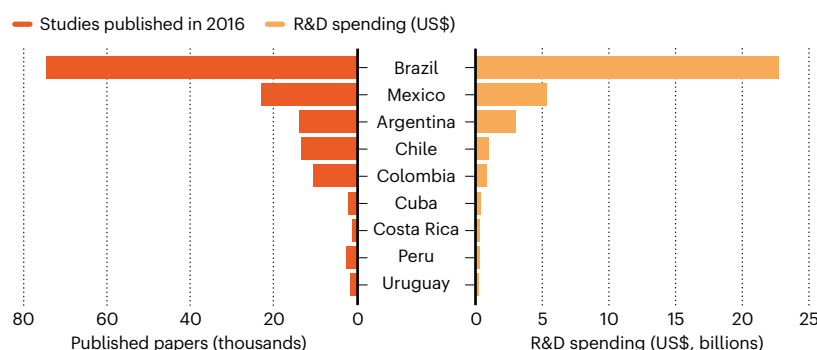
But by 2017, Orjeda realized that Peru

BY THE NUMBERS

Peru's scientific output is small compared with larger countries in South and Central America, but is increasing rapidly.

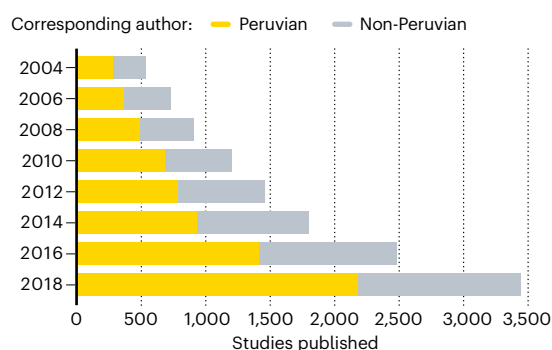
Money matters

The research and development (R&D) budgets of countries in South and Central America vary hugely, but some punch above their weight when it comes to scientific output.



Leading the way

Peru's scientific output has increased greatly since 2004 – and the share of research led by a Peruvian scientist has also crept upwards.



SOURCES: MONEY MATTERS: SCIMAGO/WORLD BANK; LEADING THE WAY: FÉLIX DE MOYA ANEGÓN



Lucas Oleniuk/Toronto Star via Getty
Researchers capture mosquitos as part of their studies of malaria in Peru.

needed a more systemic approach that addressed societal deficiencies not only in how it supports science, but also in how it teaches it. If Peru was going to develop a culture of science, Orjeda concluded, the country needed serious education reform.

“We have lousy university education,” she explains. “Peruvian policymakers don’t understand the role of universities – it is not to train professionals. The fundamental role of a university is to be a producer of knowledge, to teach students a way of thinking.”

Conservation biologist Patricia Majluf agrees that there is a deep-seated problem with the educational system. Majluf, who is vice-president of the ocean-conservation advocacy organization, Oceana Peru and former director of the Center for Environmental Sustainability at the Cayetano Heredia University, says that serious problems in higher education are hampering innovation and scientific development.

One key issue is a lack of encouragement for original thinking, says Majluf. “We’re taught in schools to regurgitate what the professor tells you. Universities haven’t really changed at all in their teaching methods,” Majluf explains. “So if you want to push society to the next stage of thinking and asking questions, we’re not really well prepared for that.”

This antiquated educational system, Majluf says, traps science in the academic world. Little is translated to society or makes an impact on

the economy, and thus on policy decisions in Peru, she says. Instead, “economics is leading every decision the government makes. Science is always in second or last place.”

Nature asked the Peruvian Ministry of Education and CONCYTEC to respond, but had not heard back by the time this article went to press.

Majluf says Peruvian scientists must learn to be effective advocates and interface with the public. If there’s one thing she learnt from her graduate work at the University of Cambridge, UK, she says, it was how to write powerful proposals, which meant advocating for oneself and one’s science. That advocacy is something she’s carried throughout her career.

Shoal food

In 2012, Majluf took a role as a minister of fisheries for the Peruvian government. This was to advocate for science being used more thoroughly to inform policy decisions. It was a change of pace from someone who, she readily admits, “went into science to not deal with people”. So moving into a political world was “completely contradictory. But it’s what you have to do when you see the planet disappearing in front of your eyes.”

One campaign Majluf brought to the ministry was the protection of Peru’s anchoveta stocks, which is one of the world’s most productive fisheries, engorged by the phytoplankton-rich Humboldt Current. By convening different

stakeholders and drafting legislative proposals informed by science, Majluf was able to help the government to set catch limits on the fish – which is exported as fishmeal for livestock feed to countries such as China, generating more than \$1.5 billion in 2018. She also helped it to pass a 2013 resolution urging Peruvians to eat the anchoveta.

With the help of Peruvian celebrity chef Gastón Acurio, the country began transforming that supply chain and reducing the impact on fish stocks. “As soon as they saw there was a real market for it, people started investing in factories and infrastructure to process anchoveta,” Majluf recalls. “Within a few years, you’d go to the supermarket and all you could see was anchoveta.”

Times have changed since, however, and those market-based mechanisms to encourage Peruvians to eat the fish have gone backwards owing to political disagreements over who has the right to catch them. Today, most anchoveta are again exported rather than eaten in Peru.

Research potential

Other efforts to bridge science and politics have lasted longer. One example is a study, published in *Nature* in May that explored rural nutrition and obesity in 200 countries, including Peru (NCD Risk Factor Collaboration *Nature* **569**, 260–264; 2019). It showed a large rise in obesity in rural areas, especially in Peruvian men. Jaime Miranda, director of the Centre of Excellence in Chronic Diseases (Cronicas) at the Cayetano Heredia University and leader of the study, says it’s one of a series of population-based studies, interventions and investigations linked to health systems that he has conducted across Peru.

“Our *Nature* paper gave us a meeting with the Ministry of Development and Social Inclusion,” says Miranda; it gave the ministry evidence to investigate something it already suspected: that Peruvians in rural areas were becoming obese.

These cases of individual scientists informing legislation and policy epitomize opportunities for impact beyond research and publications.

Carrasco-Escobar, the PhD student at the University of California, San Diego, says research has real potential to improve Peru’s economic and medical health. “There’s a saying about Peru that we’re a ‘beggar seated on a bench made of gold’. I wouldn’t be that radical, but I do think science research in Peru is still immature. We’re not aware of all the potential we have.”

Aleszu Bajak is a freelance science journalist who teaches at Northeastern University in Boston, Massachusetts.

events guide

How to make the most of conferences

Chief careers editor

David Payne

Editorial

Jack Leeming, Joanna Beckett,
Anne Haggart

Art & design

Mohamed Ashour, Denis Mallet,
Kate Duncan

Production

Kay Lewis, Ian Pope, Jason Rayment,
Nick Bruni

Marketing

Claire Jones, Alejandro Medina
Perez, Alison Price

Sales

Matt Clare, Kasia Orlowski

Director, global career services

Nils Moeller

Project management

Rebecca Jones

Editorial director

Stephen Pincock

Art director

Wojtek Urbanek

Publisher

Richard Hughes

Editor-in-chief

Magdalena Skipper

What do the Woodstock music festival and *Nature* have in common? They both celebrated notable anniversaries in 2019. The iconic music event, held in Bethel, New York, took place 50 years ago; the journal was founded 100 years earlier.

Both have also filtered into the public consciousness: the structure of DNA, published in *Nature*, is arguably as well known as Jimi Hendrix's famed performance of 'The Star-Spangled Banner' that closed the Woodstock festival.

Finally, both the festival and *Nature*'s anniversary are covered in the 2020 Events Guide.

The guide has events listings for many of the conferences available to researchers over the coming year, and aims to help the process of networking and collaboration.

Included is a comparison between scientific conferences of 150 years ago and the issues that organizers face today (see page S70). We also explore how individuals with disabilities are finding conferences suited to their needs and advocating for change in those meetings that don't do enough (S74). And we speak to one meeting organizer who hopes to host the 'Woodstock of Biology' next year. There will be walk-up songs for presenters, but no word yet on any iconic guitar performances (S76).

See go.nature.com/eventsguide for more coverage and advice on how to get the most from conferences.

We hope you enjoy reading our guide, and that you succeed in any conferences you attend in 2020.

Jack Leeming

Editor, Events Guide

Contents

S70 The past and present of conferences

What conferences were like 150 years ago — and what they're like now.

S74 Disabilities at meetings

Researchers share experiences on conference accessibility.

S76 The Woodstock of science

We meet the organizer of a conference dedicated to biological tweeters.

Nature Events Guide

The Nature Events Guide 2020, a supplement to *Nature*, is produced by Nature Research, the flagship science portfolio of Springer Nature. Nature Events is the essential reference guide to scientific events worldwide.

Nature editorial offices

The Campus,
4 Crinan Street,
London N1 9XW, UK
Tel: +44 (0)20 7833 4000
Fax: +44 (0)20 7843 4596/7

Customer services

To advertise with Nature Careers, please visit naturecareers.com or e-mail naturecareers@nature.com or feedback@nature.com. Copyright © 2019 Springer Nature Limited, part of Springer Nature. All rights reserved.



On the cover

A cluster of conference microphones. Cover image: Peter Dazeley/Getty Images Plus

CHRONICLING CONTEMPORARY CONFERENCE CULTURE

How have scientific conferences changed since *Nature* was founded 150 years ago? **By Virginia Gewin**

Participation in scientific conferences has ballooned since *Nature*'s founding in 1869. Around 24,000 attendees from 113 countries were at the 2018 American Geophysical Union's Fall Meeting in Washington DC, for example. By contrast, fewer than 2,000 tickets were sold to spectators and presenters at the 1869 meeting of the British Association for the Advancement of Science, held in Exeter, UK.

Not surprisingly, the diversity of scientists has also changed in that time, as has recognition of the environmental impact of such large gatherings. Today's conference organizers are constantly taking steps to make sure that meetings are inclusive, safe and socially responsible.

To set the scene, *Nature* spoke to a science historian who reflects on an early series of UK meetings that began almost 190 years ago. We then hear from some of today's conference organizers, who outline a range of current concerns and how to handle them – including codes of conduct, offsetting carbon emissions, recognizing gender pronouns and overcoming visa hurdles.

ALEX CSISZAR THE 1869 CONFERENCE VIBE

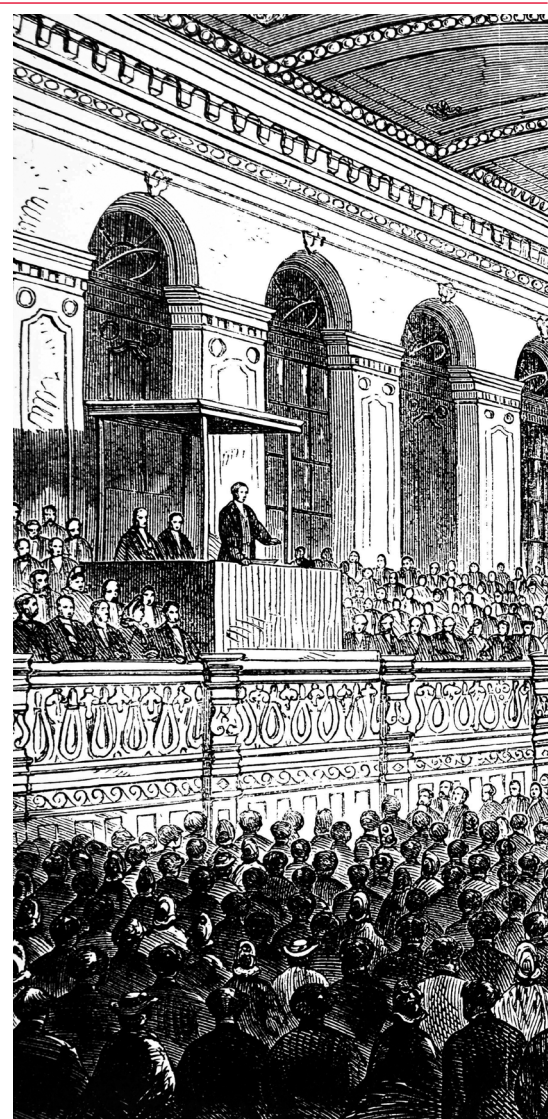
The largest regular scientific conferences in around 1869 were probably those held by the British Association for the Advancement of Science (BAAS, now the British Science Association), founded in 1831. These were modelled on a format first used by German scientists in the 1820s that had been picked up by scientific organizations. In the early to mid-nineteenth century, there was an outcry about how hard it was to make a living as a scientist, so the BAAS conference was predominantly about making science more accessible to the public. Its founders thought that if the public was interested in science, the British government might be inclined to offer more funding.

According to *Gentlemen of Science*, a 1981 book by science historians Arnold Thackray and Jack Morrell, the events hosted 1,000 or more spectators. The intention was for attendees to have a good time, forge alliances and establish research collaborations. To include as many participants as possible and circumvent travel limitations, the conference moved to a different British city each year. This was an early attempt to be inclusive – at least for the mostly white middle- and upper-class men who attended.

To help to raise the profile of British science, experts were courted to come to the meeting to discuss hot topics in science, sometimes from as far away as the United States and South America. Many spectators attended these high-profile talks (similar to today's plenary speeches), and there could be hundreds of papers presented in the accompanying sectional meetings. These were dedicated to specialist branches of science, including mathematics, geology and zoology.

Abstracts of interest were written up in newspapers and periodicals, such as the weekly literary magazine *The Athenaeum* (published from 1828 until 1921) and, later, in *Nature*. (The botanist Joseph Hooker, who worked at the Royal Botanical Gardens at Kew near London, suspected that *Nature* would fail because it was trying to do what *The Athenaeum* already did.) If you were an elite scientist and you wanted international colleagues to know about your research, you might aim to get your paper covered in *The Athenaeum* because readers often passed their copies on to others. That meant reports of your work were likely to travel overseas faster than papers in what we'd now call primary research journals. And even if a scientist didn't attend a conference in person, they might ask someone else to read out their abstract to improve the chances of it getting covered.

Officially, women were banned from the sectional meetings until 1839 because of blatant sexism, but in practice they often attended. The conferences were large affairs, and women helped to make them a success

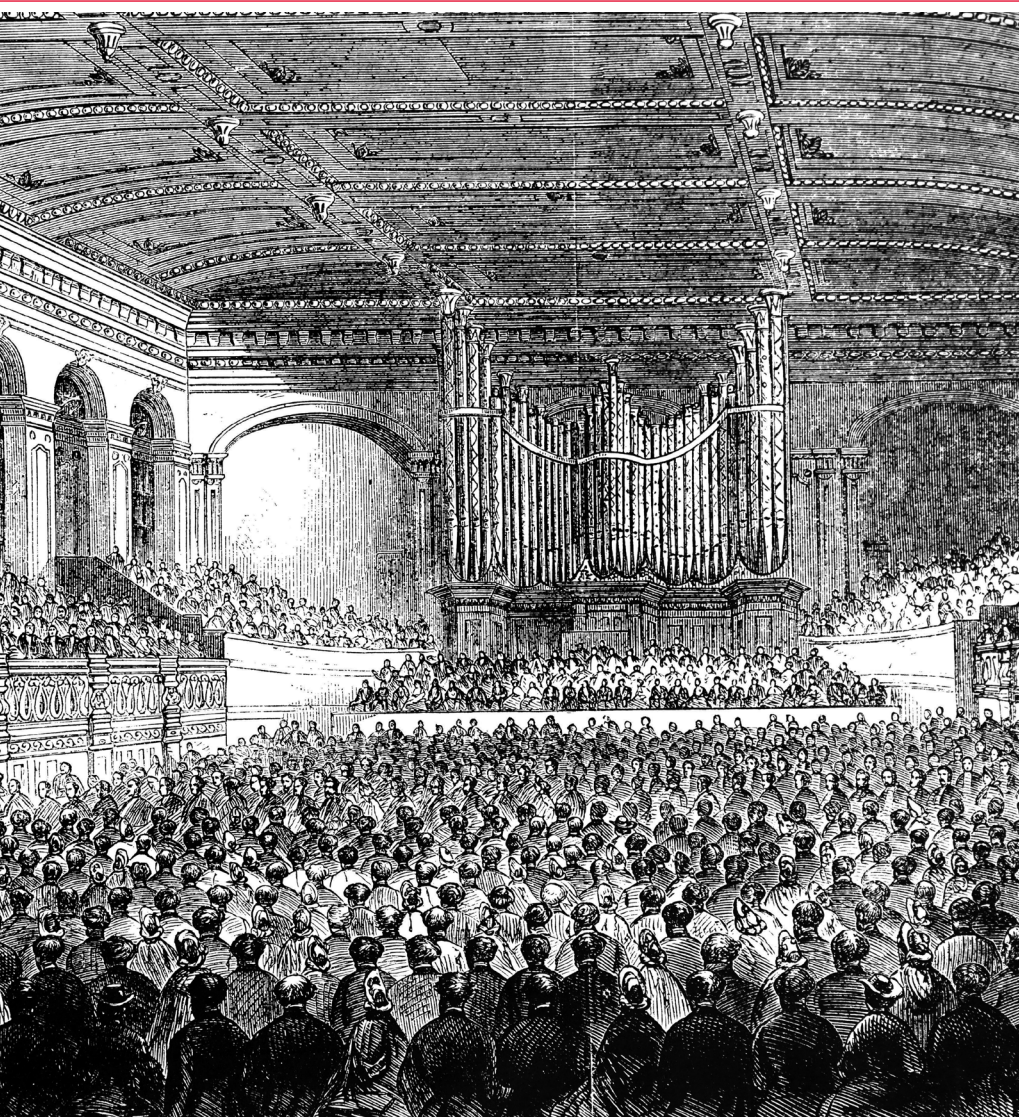


– both financially (because male members of the BAAS were allowed to buy extra tickets for women accompanying them to the meeting) and socially (because women were seen as providing cohesion for social activities

“The conferences were large affairs, and women helped to make them a success – both financially and socially.”

outside the formal sessions). Women rarely gave talks at BAAS meetings during its first years, but some did, most often in the fields of statistics and economics.

Alex Csiszar is a science historian at Harvard University in Cambridge, Massachusetts.



The British Association for the Advancement of Science held large conferences from 1831.

ANDREA CASE CODE OF CONDUCT

In early 2017, the organizers of an annual evolution meeting got together to address concerns about how to handle inappropriate behaviour at conferences – whether sexual or gender harassment, exclusion or discrimination. For our meeting that June, held jointly by the Society for the Study of Evolution, the American Society of Naturalists and the Society of Systematic Biologists, we established a formal code of conduct. Because it was a new policy, conference attendees had to agree to it when registering for the meeting. Unfortunately, I think a lot of people just clicked ‘accept’ on the form and didn’t read the policy thoroughly.

The harder part, we discovered, was what

to do if someone violated the code. And we hadn’t publicized the fact that there was an official procedure for reporting inappropriate behaviour, so one of the three reports of code-of-conduct violations at the 2017 meeting was made through Twitter.

After the meeting, we set up a special committee to help us ensure that we have robust, enforceable procedures that are well publicized. We hired a harassment-prevention consultant and created a web resource called SAFE Evolution (see go.nature.com/2e9y1sg). It is modelled on a similar initiative by the American Geophysical Union, and details types of inappropriate behaviour, procedures for reporting a violation of the code of conduct, and transparency reports from our meetings.

The following meeting, in 2018, was held jointly with the European Society for Evolutionary Biology, and its representatives hadn’t

thought about a code of conduct. Culturally, it was a bit different. Some people initially didn’t share our concerns, whereas others thought it was a great idea. Together, we agreed to have an external safety officer, who was not affiliated with our field, to receive reports of inappropriate behaviour. We always try to make sure that it doesn’t feel like we are policing attendees, but that they still feel safe. At this year’s meeting in Providence, Rhode Island, we also had 23 people dotted around the conference wearing huge badges. That meant they were identifiable to participants, and could offer assistance to those deciding whether to report a violation.

Making the code of conduct about all types of inappropriate behaviour keeps the focus on broader efforts to increase inclusivity. If you just concentrate on gender and sexual harassment, it tends to focus efforts on male harassers and female targets. We had nine reports this year. Although none was a report of clear sexual harassment, some participants said inappropriate things to others. The more often we can take reports and have conversations about inappropriate behaviour, the less likely such behaviour is to happen.

Andrea Case is executive vice-president of the Society for the Study of Evolution and an evolutionary biologist at Kent State University in Kent, Ohio.

SUSANNE BUITER GREEN THE MEETINGS

With more than 16,000 participants, the General Assembly of the European Geosciences Union (EGU) is the largest geoscience conference in Europe. As researchers who are keenly aware of environmental concerns, we realized that the flights and hotel stays for the conference, held each year in Vienna, yielded significant carbon dioxide emissions.

In 2018, for the first time, our registration process offered a voluntary contribution to offset carbon emissions. To calculate the amount to donate, a registrant could tick a box that said where they were travelling from. We raised nearly €17,000 (US\$18,800), which we gave to an anti-deforestation project in Brazil. In 2018, 32% of participants donated to the offset fund. This year it went down to 25% – a decrease due, at least in part, to the fact that many institutes now already offset their staff members’ travel emissions. In October, the EGU announced that it will offset travel emissions for every General Assembly participant at future meetings.

For us, it’s part of a bigger attempt to make

How to make the most of conferences

events guide



From 2020, the European Geosciences Union will offset emissions for assembly attendees.

our meetings greener. We have reduced paper use considerably by getting rid of the physical programme book. To reduce plastic, we have water fountains and ask attendees to bring their own bottles. Next year, we will ask people to bring their own coffee mugs.

To discourage unnecessary duplicate travel, we also recommend that attendees who have other business with conference participants book a small meeting hall for free. And we are exploring how to accommodate remote participation using web streaming.

Susanne Buiter is a programme chair at the European Geosciences Union and a solid-Earth geologist with the Geological Survey of Norway.

SARAH MILLER IDENTIFY GENDER PRONOUNS

There has been a big shift in the past five years in how we talk about gender diversity. There are many different ways to be gendered and it is important to make sure we don't misgender people. For roughly the past decade, the annual American Sociological Association (ASA) conference, our field's biggest event, has encouraged attendees to write their preferred pronouns on their badges. However, cisgender people (those whose gender identity matches

the sex they were assigned at birth) often don't participate. That creates a problem because it further alienates those who do take part. When everyone states their preferred pronouns, it's more equitable for all.

Working with the ASA, our committee helped to institutionalize pronoun selection – we

“When everyone states their preferred pronouns, it's more equitable for all.”

asked attendees to designate their preferred pronouns as part of last year's registration process, which were then printed on their badges. At the last conference, the options were limited to she/her, he/him, they/them, ze/zir or ze/hir. Ideally, it's best to include an open-ended option that allows the registrant to add any other preferred pronouns. Everyone has the ability to opt out, but it shouldn't be the default.

There are radical changes occurring in the number of people openly identifying under the trans umbrella. And younger people are more likely than older ones to identify as transgender or non-binary. The sooner conferences get on board with these shifts, the better.

Sarah Miller is the American Sociological Association's member on the Sociologists

for Trans Justice committee on promoting inclusion at conferences, and a sociologist at Boston University, Massachusetts.

RICHARD HUGANIR REMOTE PARTICIPATION

Getting visas in general is a slow, bureaucratic process. But to my knowledge, it wasn't a huge deal until the introduction of the 2017 travel ban, which denies individuals from several Muslim-majority countries entry to the United States. Around 30–40% of members of the Society for Neuroscience (SfN) are international. We have an obligation to them. The expansion of neuroscience and international brain-research initiatives make the field a more collaborative environment.

During my year-long tenure as president of the SfN from November 2017, the immigration issue became a big deal. It has also become tougher to get a visa for those in China or Mexico, for example. Members were upset that researchers from certain countries who submitted abstracts couldn't come to the meeting. Some students from places such as Iran who were travelling from other countries not affected by the ban were also prevented from getting visas.

For the first couple of years, we tried to provide letters of support for visas. Once the immigration ban went into effect, some individuals had to cancel their trips to the meeting. We decided to encourage those people who couldn't get visas to present remotely: accommodations we made under the Science Knows No Borders programme (see go.nature.com/2ycteds).

We had around ten remote presenters at this year's conference, including some who provided pre-recorded Powerpoint presentations. Remote poster presenters were encouraged to conduct an online chat to allow meeting attendees to ask questions. I've never heard of other societies doing this. Although some remote presenters expressed disappointment about the limited feedback and engagement they received, this was the programme's first year, so the SfN will learn from the experience and try to improve it next year.

Richard Huganir is a former president of the Society for Neuroscience and a neuroscientist at Johns Hopkins University in Baltimore, Maryland.

Interviews by Virginia Gewin.

Interviews have been edited for length and clarity.

EGU/FOTO PELUEGL



Astronomer Wanda Diaz-Merced uses apps that improve conference accessibility.

CONFERENCE CONUNDRUMS

Scientists share their experiences of accessible meetings. **By Emily Sohn**

Conferences can be tough for people with disabilities and chronic illnesses. Four scientists talk about their own conference challenges and positive experiences, and give advice for other researchers and event organizers.

MONKOL LEK PLAN AHEAD

I have a rare form of muscular dystrophy, and can't walk long distances. I use a wheelchair, or I just struggle. I trip easily and can't get off the ground when I do.

When people are invited to give conference talks, everyone gets so excited. My first thought is 'how am I going to get on the stage?' I also wonder if the conference venue is accessible, how far it is from the hotel and how I can sit in such a way as to not to cause a ruckus.

I'm thinking about all the things that people without a disability just don't think about. It takes the fun out of being selected to give a talk.

I can tell you about one epic fail. I was speaking at the 2019 annual meeting on rare diseases in Beijing, hosted by the Beijing Society of Rare Diseases. I was the only international speaker in that session; there were about 1,000 people in the audience. I was walking with a walking stick from a seat close to the stage, which was elevated by only 60–80 centimetres, so it was possible for me to transfer from a standing position to a chair on the stage that I had asked for. The transfer went fine, but when I got up from the chair to speak, there were too many people crowding me to help. I tripped over their feet and had to use the chair to get up off the stage floor. It was really embarrassing and, looking back, I kind of did a shit job on the talk. My hosts are always lovely, but they can forget about accessibility because it's something they don't deal with every day.

I go to three to four conferences, university visits and other meetings a year. I probably turn down three to four more because my body can't take all the travelling. I travel with my wife; she is a scientist and we have to balance her career goals, too. My colleagues are constantly at conferences, and I feel like I'm missing out on opportunities to network and build new collaborations.

My advice to people with physical disabilities is: don't be embarrassed to contact event organizers in advance. Before I went to the Biology of Genomes conference at Cold Spring Harbor in New York, I learnt there was a long walk between the conference centre and where we were staying. The organizers confirmed there was disabled parking available. That made my life so much easier.

Monkol Lek is an assistant professor at the Department of Genetics, Yale School of Medicine, New Haven, Connecticut.

WANDA DIAZ-MERCED NETWORKING CHALLENGE

Being blind, the major challenge I face at meetings is networking. Trying to locate people, knowing who is in the room, trying to approach and talk to them is challenging. I am unable to have spontaneous conversations.

I go to four or five astronomy meetings a year, and there is a lot of preparation to do. I call the hotel and conference venue and ask about accessibility, whether the hotel is easy for taxis to reach, and if they have markings on the floor so I can find my room and the breakfast place. I like to do things by myself.

There are many things I miss out on. If conference rooms are too distant from each other and there are landmarks that I am not aware of because I didn't attend an orientation, I might lose the opportunity to go to sessions. It's hard to ask anyone walking around to take me to a room: I don't know who is there or how distant they are. Bringing a companion can solve the situation, but not all astronomical observatories have the money, and I worry that asking them to pay for someone to accompany me might mean I lose opportunities in the future.

I presented at a TED conference called Dream in Vancouver, Canada, in 2016. It was a completely different experience. It had an app that let you know who would be at the conference and what they would be talking about, so I could get in touch and plan to meet with people in advance. There was a sensor in your conference badge that was always scanning and updating the app with the location of



KRISTINA FORMUZAL

Biochemist Bryan Fry, who has hearing loss, studies the venom of Komodo dragons.

people, so you knew who was in the room. The organizers also gave information about dietary requirements on the menu. I'm diabetic and I didn't have to point that out to anyone. Those accommodations made a huge difference, I felt free to be myself. I still keep in touch with 10 or 12 people that I met at that conference.

My advice to organizers is to ensure your conference allows everyone to interact.

Wanda Diaz-Merced is an independent professional astronomer.

BRYAN FRY AUDIO TECHNOLOGY

Audio at conferences is not always the best, and can be a challenge as a person with hearing loss.

I remember an awkward situation at one question-and-answer session when I was presenting, and one person who asked a question wasn't speaking very clearly. They had a thick accent, but that wasn't the problem. It was that they weren't using the microphone properly. I kept asking them to repeat the question, and they ended up thinking that I was making fun of their accent. I could see them getting visibly upset. A couple of people giggled in the audience. It led to a colossal misunderstanding. I explained to them afterwards that I had hearing loss, but I don't think they believed me.

Because of nerve damage in my ear, there's no hearing device that works for me. If someone's talking on my right side, they might as well not exist. At round-table meetings or conference dinners, if there are more than four people or if we're at a loud restaurant, I have no

hope of staying involved in the conversation. After a while, it becomes incredibly isolating. I also have bad balance that's related to hearing loss. When I'm walking through a crowd, I might bump into people, or I'll walk into a wall and someone might think that I'm drunk. I have to accept that there's going to be some misunderstanding or lack of participation and interaction. Not all disabilities are visible.

The best conference I have been to was 'Snakebite – from science to society', last year in the Netherlands, which I helped to arrange. I brought up hearing and the other organizers had already made plans for accessibility. They had a number of small loudspeakers dotted around the conference hall, so that you could be sitting at the very back and have the same level of volume directed at you as someone sitting at the front. They were also clear about microphone use. They actually stopped a couple of people and gave guidance about holding the microphone.

"Don't be embarrassed to contact event organizers in advance."

I design my slides with key bullet points accompanying images, so that you could be wearing noise-cancelling headphones and still get the essential information. Often slides are extremely pictorial, but if you can't hear what the speaker is saying, you can't understand. I don't think it's lack of caring. I think there's just a lack of awareness about these sorts of issues.

Bryan Fry is a venom researcher and

biochemist at the University of Queensland, Brisbane, Australia.

GABI SERRATO MARKS EMPATHY FOR YOURSELF

Because I have Ehlers-Danlos syndrome, which is a connective-tissue disorder, I struggle standing for more than 10 minutes, which makes dinners or receptions and other networking sessions difficult. Often there are no non-alcoholic drinks apart from water, which is frustrating. Many people with disabilities and chronic illnesses don't drink because it can exacerbate our symptoms. But it's nice to have a drink in your hand to match everyone else.

I visited one conference a while ago and there were only standing tables and finger foods, most of which I couldn't eat because I have dietary restrictions. Some of the food wasn't labelled. It felt like it wasn't designed for people with disabilities.

This year, the Geological Society of America conference in Phoenix, Arizona, was better. It had a section in the registration form that solicited suggestions to make the conference better ahead of the event. Someone reached out before the conference and asked what I needed. Having a person there to help made me feel like I belonged. That support meant that I could actually present a poster and be there; otherwise I would have declined the invitation.

When I was first diagnosed, I felt so alone and as if I couldn't be a geoscientist if I had a disability. That's part of why I try to be really vocal about it – writing opinion pieces, using social media, speaking at conferences and speaking up in person when I need accommodations. I get messages from people who say, "Oh, I've dealt with health issues for my whole career and I keep it pretty quiet, but I really appreciate that you're talking about this."

I encourage anyone who has a disability or a health issue to try to imagine if a friend came to you and said, "I feel really exhausted and I don't think I can make it to this event. Do you think that's OK?" You would probably tell them, "Of course it's OK. You should go home." Have the same empathy for yourself. You should do whatever you need to do to be able to access the conference. That's the important thing.

Gabi Serrato Marks is a PhD candidate in marine geology at the Massachusetts Institute of Technology, Cambridge.

Interviews by Emily Sohn.

Interviews have been edited for length and clarity.

Oded Rechavi Setting up the Woodstock of biology

Molecular biologist Oded Rechavi uses nematodes to investigate epigenetic inheritance at Tel Aviv University in Israel. Rechavi has organized what he calls the Woodstock of Biology, a free conference for scientists he enjoys interacting with on Twitter. It will take place on the Tel Aviv University campus from 13 to 14 February 2020, a few days before an experimental-biology conference in Eilat. At the Woodstock conference, each attendee will present a short talk outlining new, unpublished work using two slides simultaneously posted on Twitter; questions and answers will follow online.

How did you come up with the idea for this conference?

It was midnight on a Friday in June this year. I had some nice exchanges with scientists whom I know only from Twitter. The biology community on Twitter is very supportive and progressive. I said to myself that it would be nice to meet these people and form a real community instead of a virtual one. So I tweeted that I would be happy to organize a conference for these scientists to meet in person. I woke up in the morning and I saw that there were hundreds of responses. People kept tweeting, retweeting, “sign me up.” So we built a website and close to 200 people signed up.

Why is Twitter useful for science?

Twitter is amazing. You learn everything in real time, as it happens. People read your paper; they react fast. Specialists respond, and you can reply to them. It's all online and open. It's a revolution in the way that we spread information about science.

Why is the Woodstock of Biology important?

I'm hoping to reproduce that informal, fun, real-time involvement with peers and non-specialists that you find on Twitter and make it work in a conference format. The science community on Twitter is big, but it's not always clear how it overlaps with the real science community. This meeting is the first time, to my knowledge, that we're joining the two worlds, and the potential is that some of the good stuff inside the Twitter science



community will penetrate the real science world and influence it.

Why did you call it the Woodstock of Biology?

It was coined by Shai Biran — an acquaintance who works at the biotech consultancy firm MacDougall in Natick, Massachusetts. It symbolizes what the original Woodstock music festival symbolized: counterculture and freedom from social conventions, and peace and music too.

How will this conference be different from all other scientific conferences?

We want it to be collaborative and non-hierarchical, without a moderator. People who are not present at the conference can also respond online with questions. We'll have walk-up songs for each speaker, which tells us a little bit about their personalities.

How does this meeting help scientists?

When you go to some conferences, you hear things you already know. People are afraid of being scooped, so avoid speaking about completely new things that are far away from publication. But if you want conferences to stay relevant, you need to encourage speakers to talk about new stuff. I think because so many tweeters and science communicators

will be at the conference looking at the work, it doesn't give anyone else the chance to scoop work without being caught — in that way, presenting at the conference will be like publishing a preprint.

What do you hope to achieve in this conference?

I hope to get some collaborations going for myself and others. I'm hopeful that it will be a model of how conferences should be: when you do something in a fun and friendly way, it improves the chances that people will be friendlier and enjoy themselves, and not be defensive or competitive.

Which speakers have signed up?

One keynote speaker is Dan Shechtman at the Technion, Israel's Institute of Technology in Haifa, who won the 2011 Nobel Prize in Chemistry. Another is neurobiologist Piali Sengupta at Brandeis University in Waltham, Massachusetts. The third is systems biologist Uri Alon at the Weizmann Institute of Science in Rehovot, Israel. We will have PhD students talking alongside a Nobel prizewinner, and will cover a range of disciplines: from ecology and molecular biology to neuroscience and philosophy of science.

What are some of the advantages for early-career researchers attending this conference?

In a really practical way they have an opportunity to talk about their research, as equals, next to really prominent scientists from different fields. I think it will help them to find positions, get connected, help them to tweet their stories and increase their visibility.

What advice would you give to others organizing similar conferences?

We should let young people talk, and not just invite the same people over and over. I think that the format we are inventing could also be mimicked. I would be happy to have walk-up songs at every conference. Making conferences fun is something I think will help science.

Interview by Josie Glausiusz

This interview has been edited for length and clarity.